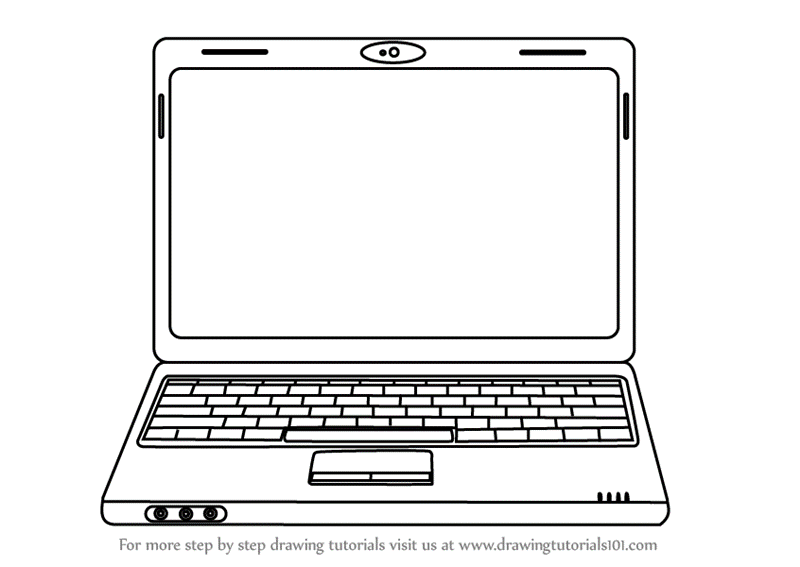
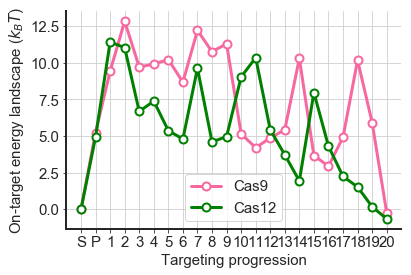
|  |
| --- |
| A model to predict off-target binding of CRISPR-Cas systems |

|  |
| --- |
| **Diewertje Dekker**  **6-1-2019** |



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   
https://www.drawingtutorials101.com/how-to-draw-a-laptop-1

Abstract

The CRISPR-Cas system is used as a gene editing tool in many fields like biotechnology and clinical medicine, ~~but also to develop experimental setups~~. One occurring problem is that next to the target sites, off-target sites will also get targeted, resulting in that it is unknown at which places the DNA will get cut. The group of Martin Depken at TU Delft is building a kinetic model for the CRISPR-Cas system in the hope to be able to predict the probability of binding to (and cutting) an off-target sequence. They have succeeded in training this model on the data of Boyle et al. and predicted a different dataset successfully.

For this thesis I trained the model on a different type of dataset (from Finkelstein et al.), obtained from a different experimental setup, which is executed for multiple different Cas proteins. We discovered that this type of data does contain enough information to train the model and get a good fit to the data. However, we also discovered 2 properties of the data, due to which slight changes in the setup can make the experiments more efficient. Firstly the assumption made by the experimentalists, of reaching equilibrium after 10 minutes, is not met for Cas9. We would advise … Secondly, the dynamics of Cas9 can be described by the occupancy at 3 concentrations instead of 9, which can shorten the experiment a lot.

Since the dynamics of Cas9 can be described by the data in the same extent as we saw with the data of Boyle et al. we moved on to a different protein: Cas12. …

Further recommendations…..

Table of Contents

[1 Introduction to CRISPR: 1](#_Toc11673073)

[1.1 The biology of CRISPR-Cas 1](#_Toc11673074)

[1.2 Cas9 and Cas12 2](#_Toc11673075)

[1.3 CRISPR-Cas as a gene-editing tool 2](#_Toc11673076)

[1.4 The research of the Martin Depken group 3](#_Toc11673077)

[1.5 The focus of this thesis 4](#_Toc11673078)

[2 Explanation of the model 5](#_Toc11673079)

[2.1 The general concept 5](#_Toc11673080)

[2.2 The kinetics 5](#_Toc11673081)

[2.3 Enforcing a binding equilibrium 7](#_Toc11673082)

[2.4 Other assumptions 7](#_Toc11673083)

[3 Description of the different experiments 8](#_Toc11673084)

[3.1 Finkelstein et al.: 8](#_Toc11673085)

[3.1.1 Processing of the data from Finkelstein et al. 9](#_Toc11673086)

[3.1.2 Derivation of the Hill equation 10](#_Toc11673087)

[3.2 Boyle et al.: 11](#_Toc11673088)

[4 What we actually fit 12](#_Toc11673089)

[4.1 Simulated Annealing 12](#_Toc11673090)

[4.2 Selection of the fits 14](#_Toc11673091)

[5 Discovered properties of the data by fitting 15](#_Toc11673092)

[5.1 The usage of less concentration points 16](#_Toc11673093)

[5.2 Assuming equilibrium 17](#_Toc11673094)

[5.3 Comparison to Boyle 19](#_Toc11673095)

[5.3.1 Predicting other datasets with the obtained parameters 20](#_Toc11673096)

[6 Derived properties of Cas12 22](#_Toc11673097)

[6.1 Fits based on 3 concentration points 22](#_Toc11673098)

[6.2 Assumption of equilibrium 23](#_Toc11673099)

[6.3 A closer look at the on-target energy landscape 24](#_Toc11673100)

[6.4 Comparison between Cas12 and Cas9 24](#_Toc11673101)

[7 Discussion and outlook 26](#_Toc11673102)

[8 Conclusion 27](#_Toc11673103)

[9 Acknowledgements 28](#_Toc11673104)

[10 Bibliography 30](#_Toc11673105)

1. Introduction to CRISPR:

A close up of text on a white background

Description automatically generatedClustered Regularly Interspaced Palindromic Repeats combined with a CRISPR-associated operon, or better known as CRISPR-Cas, is a well-known biological system. Especially for the ones working or studying in the field of bio-nanoscience, but also for the general public. 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 real born babies [1]. Although most scientists share the opinion that CRISPR-Cas should not be applied to human germlines until it has been extensively tested [1], 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 [2] and phylogenetic classification of bacterial strains [3]. 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 [4]. In other experimental set-ups CRISPR-Cas9 is for example used to fluorescently tag specific DNA loci, which is a powerful life-cell-imaging alternative to DNA-FISH [5]. Moreover, Cas9 can by injected into fertilized zygotes to generate transgenic animal models [6,7], or it can be used for genomic screens by introducing loss-of-function mutations in human cells [8,9]. An example of a medical application is that it has been used to create antibiotics whose spectrum of activity is chosen by design [10].

# 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. 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 : The Lytic cycle [34]

So 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 [11]. 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 [12]. In front of this CRISPR array is a series of genes that drive the three phases of immunity [11]:

(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. This is done once for every new virus, only the first time it enters the bacterium. This first time the viral DNA is recognised by its protospacer adjacent motif (PAM). This PAM region is used later as well, now in order to distinguish between viral DNA and the CRISPR-array. They both include the same DNA sequence, but the PAM region is only present in the viral DNA, not in the CRISPR array. The interference machinery is only able to bind to the PAM and therefore it is prevented that it will cut 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, which 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. How the DNA is cut depends on the Cas protein. Cas9 generates a double-stranded break 3 bp upstream of the PAM [13].

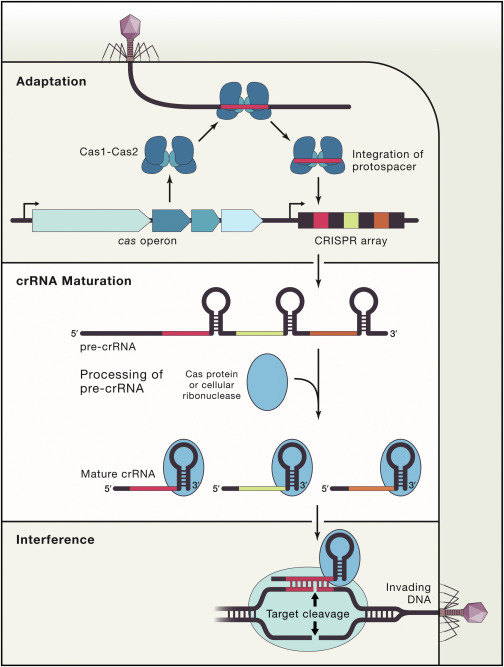


Figure : The three stages of CRISPR immunity [11]

# Cas9 and Cas12

… What is already know about them …

# CRISPR-Cas as a gene-editing tool

A series of studies [14–18] led to the realization that induced DNA double-strand breaks (DSBs) can stimulate genome editing through homologous recombination events [19]. This inducing of DSBs is, as we described in section 1.1, exactly what we can achieve with CRISPR-Cas.

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 different DNA sequence 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.

As we said before, CRISPR-Cas was originally a defence system against viruses in bacteria. From an evolutionary point of view, it would not be convenient to specifically target only one sequence, since in that case the defence mechanism would be 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 an 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 and could be catastrophic when we use CRISPR-Cas for gene-editing of the human germline [1].

Therefore, we want to understand this off-target binding better. Studies from Fu et al. [20], Hsu et al. [21], Mali et al. [22], and Pattanayak et al. [23], 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. [24] and Finkelstein et al. [25] that we will use for this thesis. Hsu [21] and Pattanayak [23] also showed that mismatches appear to be better tolerated at high concentrations of Cas9. On top of that, Wu et al. [26] 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. [21] and a model from Stemmer et al. [27], use empirically determined scoring algorithms to quantify off-target cleavage. Doench et al. [28] 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. [29] show their model that is based on the free-energy scheme of Cas9. They hope this model will be transferable to other Cas proteins, since it is based on solely physics. However, no generally accepted model exists so far.

# 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. [29] (for a comparison: see discussion (section 7…)). 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.

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 [30]. This model is trained on the experimental data of Boyle et al. [24] 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.

# 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. [25] 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. [24]. 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.

We expect this model to be applicable to other Cas proteins since it is based on pure kinetics and the only information it takes in is the guide RNA, nothing about the protein itself. 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.

So the main question for this thesis is:

“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 to Cas12 and see which properties we can determine about this protein (Chapter 6). We will conclude this thesis with a discussion and an outlook to further developments for the model.

1. Explanation of the model

The off-target activity of CRISPR-Cas forms a problem for using it as a tool to modify DNA. The Martin Depken group at TU Delft have started building a model, to gain understanding of the targeting rules, by kinetically modelling the physics of guide-target hybrid formation [30]. The 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 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 take a look into the assumptions behind this model and try to justify them as far as possible.

# The general concept

Since our main goal is to understand the off-targeting activity of CRISPR-Cas9~~, the model starts off with only already known facts and we try to build from there, according to the laws of statistical physics~~. From experiments [31] 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 one would be defined as the state where CRISPR-Cas9 is in solution, the second as CRISPR-Cas9 bound to PAM, 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 have to go through all the states, until the last one , from which you cannot come back anymore.

A picture containing text

Description automatically generated

Figure : A graphical representation of the model

# The kinetics

To understand the mathematics behind modelling these 23 states with their rates, let us take a 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 how easy you can go to the adjacent state. In general terms, it is most likely to go to the adjacent state with the lowest free energy barrier. The free energy is defined as: . E is the internal energy and S the conformational entropy at a fixed temperature T.

**A)**A close up of a map

Description automatically generated

**B)**A close up of a map

Description automatically generated

Figure : A) A model of crispr-cas with 3 states (in solution, bound and cleaved) and the switching rates inbetween.   
B) The free energy landscape where the states are depicted as minima [30]

We use kinetic modelling to describe the dynamics of the CRISPR-Cas system. The assumption is that the time for any single reaction to get completed () is exponentially distributed. This assumption is justified, since this is the distribution for a Poisson point process/stochastic process, which describes random events. Using this assumption, the probability density for the reaction time of every individual reaction (switching to an adjacent state), can be described as:

|  |  |  |
| --- | --- | --- |
|  |  | () |

In this equation, is the probability to complete the switching to a next state at time and 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 from state 1 to state 2. We can derive the chance to be in state 2 () from :

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  | () |

In these equations, the reaction rate is related to the free-energy barrier that has to be overcome for the reaction. This relation is described by the Arrhenius equation (4). This also shows that the lower the energy barrier, the larger the rate, so the higher the change is to proceed in that direction.

|  |  |  |
| --- | --- | --- |
|  |  | () |

If we now move back to the system with 3 states as depicted in Figure 4, we can use this same principle for the probability of being in one of those states. 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 [32]. Doing this for all the states gives the following set of differential equations:

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  | () |
|  |  | () |

This can be rewritten in a matrix-vector form ():

|  |  |  |
| --- | --- | --- |
|  |  | () |

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 . 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.

# Enforcing a binding equilibrium

For this thesis we work with dCas9 and dCas12, which lost its ability to cut and therefore we ignore the last state. This results in ~~the fact that we get~~ a closed system in which a binding equilibrium will form between the unbound and bound states as . The goal is that in this binding equilibrium the occupancy will start to follow the Boltzmann distribution as shown in equation (9). We want to achieve this goal, since this allows us to relate the energy of a specific state to the probability of being in that state. In other words: to relate the measurements (which measures the occupancy) to the energy landscape that we want to form.

|  |  |  |
| --- | --- | --- |
|  |  | () |

Here Z is the systems partition function and is the free energy for that state.

In order to get this binding equilibrium for (it might never be reached during the whole experiment), we impose a detailed balance condition as stated in equation (12) . This is obtained by saying that the occupancy over time should not change if there is a binding equilibrium.

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  | () |

To generalise this detailed balance condition to every state in our 23 state model with adjacent states and :

|  |  |  |
| --- | --- | --- |
|  |  | () |

# Other assumptions

Next to assuming a binding equilibrium for , there are 2 other assumptions we use for our model. Firstly, the rate from solution to PAM is linearly dependent on the concentration of CRISPR-Cas9, which is actually a general rule of reaction rates in chemistry. Secondly, all the internal forward rates are equal. The backward rate is determined by the relation as described in equation (12). This second assumption is made with the idea that going from solution to bound to PAM takes much longer than switching between all the other states, that stating a difference between all those rates will not add any information to the system. ………???

1. Description of the different experiments

The kinetic model of CRISPR-Cas from Klein et al. [30] is trained on the experimental data of Boyle et al. [24]. This data was used since the data itself already revealed certain patterns, indicating that features can actually 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.

Now we want to be able to train the model on the data of Finkelstein et al. [25] in order to be able to switch to other Cas proteins. This data only contains the occupancy after 10 minutes for different concentrations and therefore does not include any information about time(constants), while the data of Boyle et al. did. For this thesis we want to enable to model to train on Finkelstein’s data as well. In order to do so we need to understand the measured quantities and their data processing, since we want to exactly copy this during our fit.

# Finkelstein et al.:

Finkelstein et al. [25] developed a chip-hybridized association mapping platform (CHAMP) that combines sequencing with mapping the interaction between proteins and ~107 unique DNA clusters, equally spaced over the chip (See Figure 5). All of 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.

A screenshot of a cell phone

Description automatically generated

Figure : Overview of the CHAMP workflow [25]

The data they produced for Cas9 is not published yet. They measured the fluorescence for a library of DNA strands/clusters (the same as Boyle et al.) with all possible single and double mismatch configurations. They measured this after 10 minutes, for 9 different concentrations of Cas9 (0 nM, 0.1 nM, 0.3 nM, 1 nM, 3 nM, 10 nM, 30 nM, 100 nM, 300 nM).

The following should be taken into account:

* The measurements are done after 10 minutes. Finkelstein et al. assumed that this was enough time to reach equilibrium. But there was no confirmation or justification of this assumption. If equilibrium is not reached yet, we will only be able to say something about the first part of the landscape. The second part will simply be undefined because CRISPR-Cas did not have enough time to reach these later nucleotides.  
  In section 5.2 we will discuss if equilibrium is reached or not.

## Processing of the data from Finkelstein et al.

The data Finkelstein et al. measured is the fluorescence intensity~~, which they process such that they obtain the Apparent Binding Activity (ABA). Since we need to do exactly the same during the fit as they did, we need to understand their way of processing the data.~~

From the fluorescence intensity one can calculate the occupancy as described in Equation (13). Here is minimum of the measured intensities, which is included to normalise . ­­ is the intensity from the on-target at 300 nM, thereby setting the on-target occupancy to one.

|  |  |  |
| --- | --- | --- |
|  |  | () |

From the one can calculate the by fitting the resulting curve (of VS concentration) to the Hill-curve as depicted in equation (14). 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 model, see the derivation in section 103.1.2.The then reflects the concentration at which . (see Figure 6)

|  |  |  |
| --- | --- | --- |
|  |  | () |

From the the Apparent Binding Affinity (ABA) can be calculated. This is the difference in apparent (free energy) between the negative control sequence (a random DNA sequence that shares no complementarity with the RNA guide of the CRISPR) and the sequence of interest. Note the used convention: lower ABA values indicate stronger binding.

|  |  |  |
| --- | --- | --- |
|  |  | () |

The ABA gives the amount of free energy for each DNA strand upon binding. We can link this ABA value to the free energy landscape of CRISPR-Cas9 that we want to form, by using the mismatch information of every DNA strand and combine this with the measured/predicted ABA.

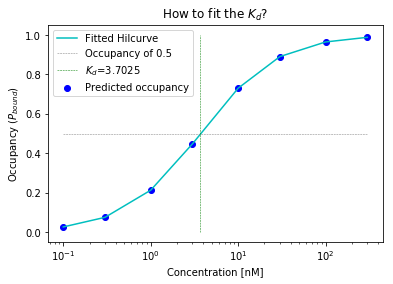


Figure : Based on the parameters the occupancy is predicted, through which the Hillcurve is fitted. at the concentration where the occupancy is equal to 0.5, the is found.

## 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: bound to solution, PAM and the 20 nucleotides (N). Again we enforce a binding equilibrium for (equation (16)), to make sure that the Boltzmann distribution holds and we end up in equilibrium at . So in using the Boltzmann distribution we assume equilibrium for the Hill-equation.

|  |  |  |
| --- | --- | --- |
|  |  | () |

|  |  |  |
| --- | --- | --- |
|  |  | () |

Since the rate from solution to PAM is dependent on the concentration, such that if you higher the concentration, the rate increases, we can state that . Using this and equation (16), we get the following expression for the energy at state PAM for every concentration compared to the reference concentration :

|  |  |  |
| --- | --- | --- |
|  |  | () |

Generalising gives the following expression for the energy at any given state n:

|  |  |  |
| --- | --- | --- |
|  |  | () |

Plugging this back into our equation for and rewriting gives the following expression:

|  |  |  |
| --- | --- | --- |
|  |  | () |

Here we take:

* since our reference concentration in the experiments

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | () |

Which leaves us with the Hill-equation:

|  |  |  |
| --- | --- | --- |
|  |  | () |

# 

# Boyle et al.:

To show the difference with the experiment of Boyle et al. [24] we will also take a short look at their experiment and data processing. Boyle et al. used a different high-throughput sequencing flow cell for their experiment, but they did in principle the same with the same library of DNA strands. The difference is that they measured the fluorescence for every strand at 3 different timepoints (500 s,1000 s, 1500 s) at a fixed concentration of 10 nM of Cas9. The fluorescence intensity for the on-target was measured after 12 hours, to be sure that equilibrium was reached. ~~Since the data they obtained from their experiment was different, their data processing differs as well.~~ To obtain the effective on- and off-rates, a straight line was fitted through the occupancy over time and forced to go through origin. ~~For the on-rate, they measured fluorescence at 3 timepoints directly after adding the RNP complexes to the flow cell.~~ This florescent signal was after the experiment converted into the occupancy in the same way as was explained for Finkelstein et al. (equation (13)). This increases linear in the beginning, but then will start to follow an exponential that gradually flattens out to reach equilibrium. The 3 measured timepoints are assumed to still be in the linear domain and therefore reflect the on-rate. (Note that their measured effective on-rate is not the same as our rate from solution to PAM, it is the change in the occupancy.) To measure the off-rate, they first incubated the flow cell 12 h overnight, to be sure that equilibrium was reached. Then they washed the flow cell with a dCas9-free buffer and measured again at 3 timepoints with 500 s in between. Here you see the same phenomena occurring, due to which fitting the straight line gives the off-rate.

In short, the datasets from the experiments from Boyle et al. [24] contained the occupancy, on-rate and off-rate of dCas9. But we should think of the following notions about the assumptions they used:

* On- and off-rate: dCas9 strand invasion behaviour is not expected to obey a simple two-state binding dynamics (which underlines the use of linearization to get (effective) association rates), since they would also get influenced by the amount and positions of the mismatches. Earlier mismatches would result in a higher off-rate or an lower on-rate.
* It is assumed that equilibrium is reached after 12 hours. Really justifying this assumption is not possible, but it sounds reasonable. 12 hours should at least be enough to go through all the binding states, because otherwise it would mean that a bacteria cannot defend itself within 12 hours against a virus.
* See comment

1. 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: the free energy needed for binding of CRISPR to the PAM region

2-21: the free energy needed for the binding of every next nucleotide position (if matches)

22-41: the penalty in free energy for every mismatch at that position (these are added to )

42: the rate for binding to the PAM region

43: the internal forward rates

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, I try to redo the experiment that Finkelstein has performed with our model. We start by choosing a random set of parameters, with which we first build the rate matrix (M). This is done for every mismatch configuration present in the library. For each configuration the free energy landscape is built by using the values and the information where the matches or mismatches are present. This free energy landscape describes the for every position at the DNA strand, so for every state in our model. By using the relation in equation (12) we can therefore calculate the backward rates from these and the randomly chosen forward rate. With both of these known we can build the rate matrix (see equation (8)).

From the rate matrix we calculate the ABA for different concentrations, as was described in sections 2.2 and 3.1.1. This can then be compared with the experimental data by taking the chi-squared value (Equation (23)), 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 chi-squared values for every strand are added and this gives an overall measure for how well these parameters describe the data.

|  |  |  |
| --- | --- | --- |
|  |  | () |

# Simulated Annealing

The process described in the section before, is done over and over again, in a loop to find the parameters that describe the data the best. This process is called Simulated Annealing, which is a form of an optimization algorithm, as was described by Kirkpatrick et al. in 1983 [33].

This algorithm is inspired by principles of statistical mechanics. Statistical mechanics applies probability theory to a system, consisting of a large number of 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. Using the common knowledge that nature always tries to use the least amount of energy, this is the state that will be occupied at a low temperature. Linking this back to our optimization problem, this would be the fit with the lowest error (in our case the chi-squared).

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 7). To explain this further we can compare this with annealing 2 metal objects together. You want them to anneal without any tears in the materials. To achieve this goal, they first melt the metal after which they cool it down very slowly, so that all the molecules in the object have approximately the same energy level. If you would cool it down rather quickly, some molecules would cool down quicker than others, which causes energy differences between the molecules, resulting in tears in the final material.

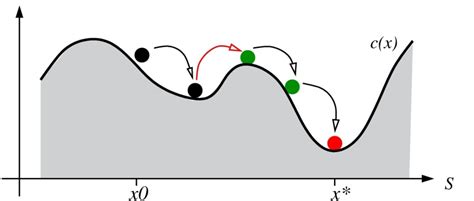
[](https://images.rapgenius.com/2706b9820eb743f64cb7da7226262e80.1000x426x1.jpg)

Figure : 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.  
(http://freesourcecode.net/matlabprojects/57563/simulated-annealing-matlab-code#.XOeodfZuI2w)

However, 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.

|  |  |  |
| --- | --- | --- |
|  |  | () |

C is the cost value of the new or old solution. In our case this is the chi-squared value of the new and old parameter set. 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 minima.

The simulated annealing protocol consists of the following steps:

1. Define some cost function to be minimised

*For us this is the function*

1. 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.*

1. 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.*

1. Accept or reject the new parameters based on equation …
2. Repeat step 3 and 4 a number of times

*In our case 1000x*

1. 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 between temperature cycles is lower then 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. Therefore the user of this algorithm should take a few considerations into account:

(Our choices for these variables are already described above).

* Initial temperature should be high enough, otherwise you get stuck in a local minimum.
* How do you change the parameters? Steps that are too big can jump over your minima, with steps that are too small will never escape your local well.
* How quickly do you want to cool down? So which cooling rate is appropriate and do you cool down every iteration? If you cool down too quickly you get stuck in your local minima, as said before. But if you cool down too slowly you can just go up and down and explore all the possible parameters without any consequence. So it will take an unnecessary amount of time to fit. Also the danger exists, that when the temperature is cooled down enough, you are at the completely wrong well, far away from the well with your global minimum.
* Which stop condition is appropriate? If it can go on for too long it can jump out of your global minimum and ends up in a local one.

# Selection of the fits

In the sections before we described what we fit ~~and how we~~ using simulated annealing ~~to achieve this~~. This whole process is done multiple times, since it is still possible to end up in a local minimum instead of the global one. Therefore we need some kind of measure to tell which of the fits describes the data the best, which is the in our case. During the fit we are minimising this , in other words, we say that the derivative of the -function should be equal to zero.

|  |  |  |
| --- | --- | --- |
|  |  | () |

In this equation is the ABA value and is the measurement error from the experiment. This formula is also known as the weighted average, which is therefore what we will use to select the best fits. Using those weights we calculate the weighted average of the data, which we assume to be the true data for our fit. Therefore we will select all the fits that have the most similarity with this weighted average of the data. The threshold is chosen such that we can see points of similarity in the energy landscape and have a ‘high’ efficiency.

1. Discovered properties of the data by fitting

The first set of fits performed used all the data that was available from Finkelstein et al. [25]. This means the for all double and single mismatches at (almost) every possible mismatch positions, for 8 concentration points, with the uncertainty in the measurement (we included the 0 nM measurement 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, but the different fits did not have a lot of agreement as can be seen in Figure 8. Later in the process we discovered that this was due to the fact that we did not put a lower bound on the (the mismatch penalty), resulting in the possibility of a negative which indicates an energy gain, which does not make sense for a mismatch penalty.

However, this was not the most urgent problem to this fit. This fit took between 40 and 60 hours to finish, while the fit on the data of Boyle et al. only needs 14 hours. Both fits were performed using 20 cores on 1 node of the HPC05 cluster of TU Delft. This difference in time is caused by the amount of points we fit, for Boyle et al. [24] this were only 3 timepoints, while for Finkelstein we have 8 concentration points. 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:* 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.
* *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 curve.
* *Assume equilibrium*: Right now the most time expensive step for this fit is the calculation of the rate-matrix. Using the assumption of an equilibrium we do not need to use this matrix anymore to calculate the . This can simply be done by adding the Boltzmann weights for every bound state:

|  |  |  |
| --- | --- | --- |
|  |  | () |

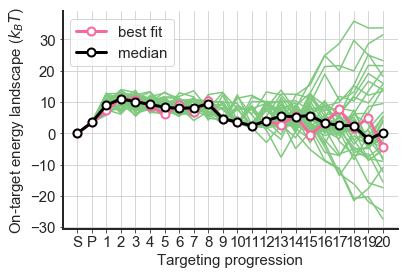


Figure : The first run of fits to the data of Finkelstein et al.. 50 fits are projected, showing a lot of diviation in the on-target Free energy landscape

# Using less concentration points

Using less concentration points would mean that you have to do less calculations and therefore make the code faster. Now the question remains which and how many points you need to still make a good prediction for the . In order to answer this question we looked back at the unprocessed data from Finkelstein et al. from which we can calculate the occupancy for every concentration. These values for the occupancy were then used in the Hill-equation to find the and calculate the as was explained in section 3.1.1. For our previous fit this was only done with all the 8 concentration points, but now we will test if we would have goten the same had we used less concentration points by evaluating the error and the correlation coefficient between the value of all concentration points and the with less points.

~~The least amount of points would be preferred, since this results in the least amount of calculations~~. Therefore we started by looking at every combination of 2 concentration points. For these we looked at the ones with the lowest error (see Figure 9A), which in this case remained at least 0.100232. Looking at the results for the combinations of 3 concentration points gives an lower error for the concentrations [1 nM, 30 nM, 100 nM] (See Figure 9B). This only gives an error of approximately 7.7% for all the strands that were tested (7395 in total) while we are using only 3 concentration points, or an error of 10% using 2 points.

**A)** A screenshot of a cell phone

Description automatically generated **B)** A screenshot of a cell phone

Description automatically generated

Figure : A) The selected concentration combinations (Error<0.2) for having 2 concentration points. B) The selected concentration combinations (Error<0.1) for having 3 concentration points.

Trying a fit with only the 2 concentration points [10,100] results in an average fitting-time of 36.58 hours. Trying the one with 3 concentration points [1,30,100] 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 ~~in the fact that~~ this now ~~becomes~~ 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 perform a bit less (See Figure 10) but do still get the overall pattern. This indicates that we can indeed use 3 concentration points for our fit.

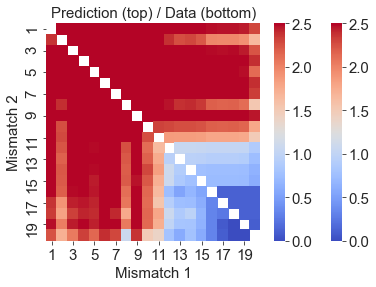
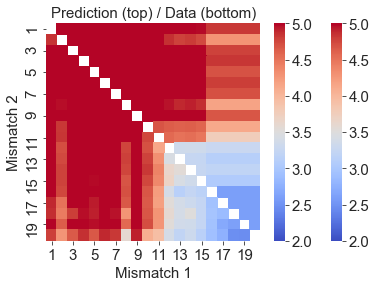
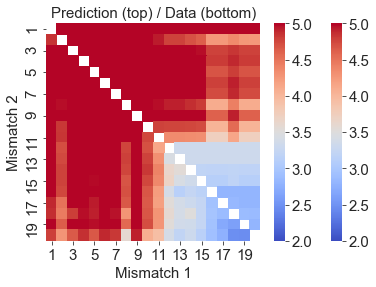
**A)****B)****C)**

Figure : The double mismatch heatmap for: A) 8 concnetration points (ABA), b) 2 concentration points (ABA),   
c) 3 concentration points (ABA)

# Assuming equilibrium

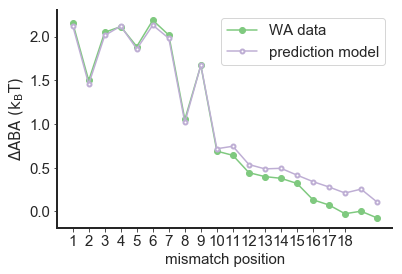
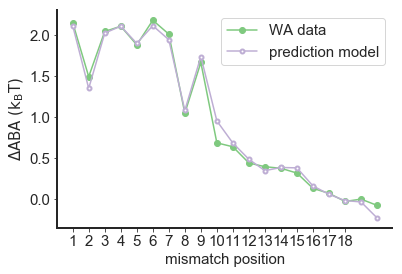
Finkelstein et al. fit their experimental data to the Hill-equation, which would actually only hold in equilibrium. Since we were mimicking their experiment with our fit, we did exactly the same: calculating (measuring) the at different concentration points and fit this to the Hill-curve.

In the derivation of the Hill-equation we saw that is equal to the sum of the Boltzmann factors (see equation (21)). Therefore, if equilibrium is actually reached, we can substitute this in the equation for ABA (15) to obtain equation (27).

|  |  |  |
| --- | --- | --- |
|  |  | () |

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 (8)) and the fitting to the Hill-curve (equation (14)), which will make the fit a lot faster.

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). Also, it fits the mismatch penalties approximately the same: values between 2 and 4 untill position 10, from there on there is a lot more deviation amongts fits. However, there is much more difference in the fitted values and the rates are not fitted at all in the fit with assumed equilibrium (since you do not use those anymore in the calculation of the ABA). Due to these differences we get a different height of the on-target energy landscape as can be seen in Figure 12.

**A)** **B)**

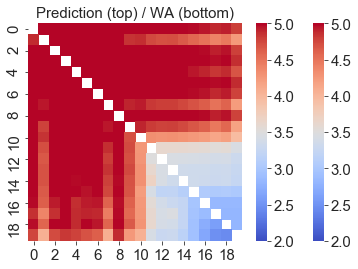
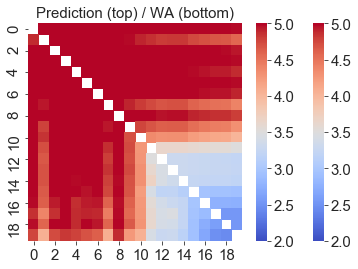
**C)** **D)**

Figure : The Single mismatch graphs (A&B) and double mismatch heatmaps (C&D) for the normal fit with 3 concentration points (A&C) and the fit where we assumed equilibrium with 3 concentration points (B&D)

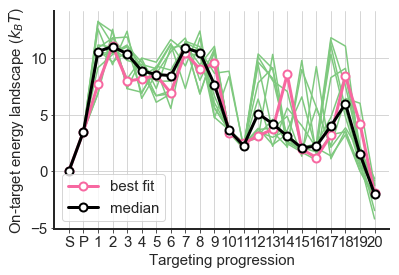
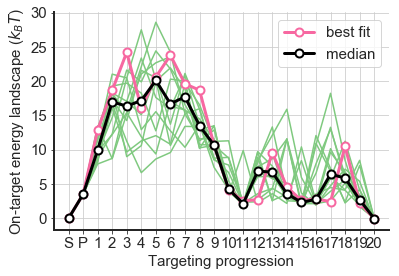
**A)** **B)**

Figure 12: The ontarget energylandschape for the fits selected on similarity with the weighted average of the data. A) has a threshold of 0.018. B) Has a threshold of 0.022.

The on-target energy landscape shows us that the fit assuming equilibrium keeps the overall general shape of the landscape and that the fits still agree on base pair position 11, which indicates the end of the seed region. However, this fit seems to just throw the landscape up 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. So apparently we need much higher rates (so throw the landscape up) to be able to reach equilibrium in 10 minutes. This already implies that there is no equilibrium reached yet during the experiment, because if it would have, there would be no need to higher the rates.

To check if there is an equilibrium after 10 min, we calculated the ABA in the two different ways, using the parameters that we got out of the fit that mimics the experiment.

1. Use the parameters to calculate the rate matrix from which we calculate that we use to fit the Hill-curve to find the value, with which we calculate ABA.
2. Use the parameters to find the energy landscape, that we use to calculate ABA with equation (26)

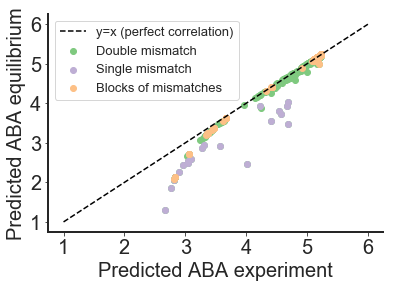
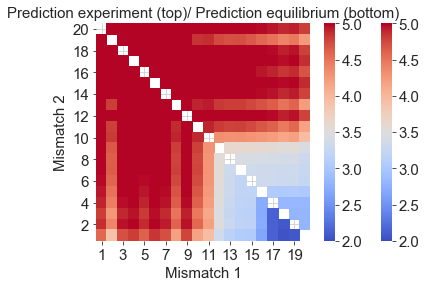
**A)** **B)**

Figure : A) The correlation between the predicted values of the ABA based on the experimental calculation or assumed equilibrium. B) The prediction for the double mismatches using both calculations

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 on the end of the sequence. 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 the off-target strands will not go through the whole landscape, it will get stopped by an energy barrier and goes back to solution. Therefore it has less states and needs less time to equilibrate. While the on-target has to equilibrate over 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
* Prediction equilibrium

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. It can be observed that the fit needs higher rates to enforce equilibrium and still be in agreement with this dataset. This indicates that equilibrium was not reached after 10 minutes.

# 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 itself. For this agreement of the fits at nucleotide 10/11 we can say that this is where the seed region ends for Cas9.

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, resulting in the fact that Cas9 simply did not have the time to explore this part of the landscape (it did not bind to those nucleotides yet). Secondly, it can be observed that the fit on the data of Finkelstein clearly suggests 2 bumps 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 actually 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 5.3.1.

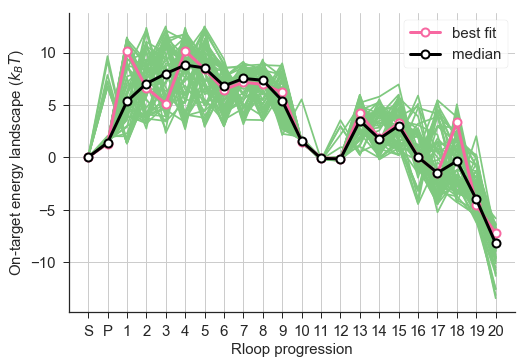
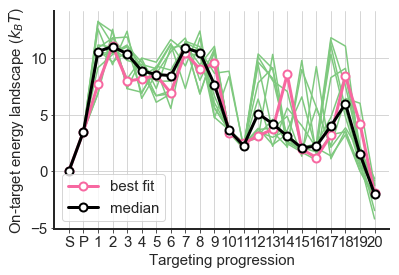
**A)**  **B)**

Figure : The on-target energy landscape A) Based on the parameters fitted on the data of Boyle et al. B) Based on the parameters fitted on the data of Finkelstein et al.

In general we can say that the on-target energy landscape shows the same characteristic as the one from the fit on Boyle. However, we observed that the parameter values do differ. The most important difference is that of . For the fit on the data of Boyle we got that 1.31, while for Finkelstein we get a value of 5.24 (which is consistent over all the fits). A possible explanation will be given in section 5.3.1.

## 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. ~~We have 2 different datasets from Finkelstein, one with only~~ 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 mismatches places consecutively (‘block mismatches’). With the parameters of Boyle we could accurately predict the block mismatches and double mismatches, but the prediction was less for the single mismatches (Figure 14).

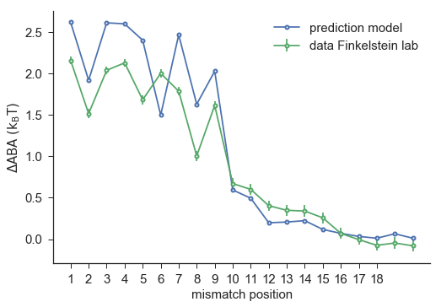
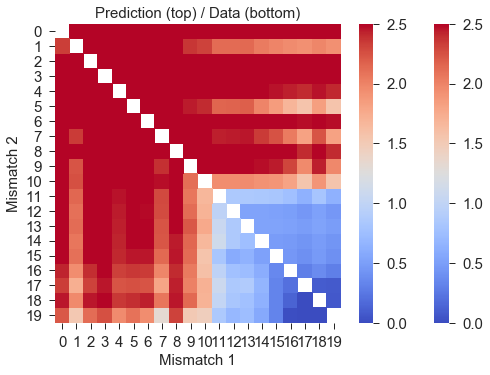
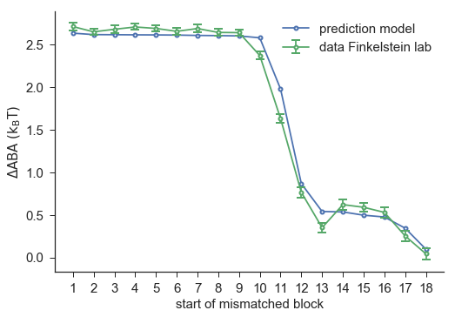
**A)****B)**   
**C)**

Figure : 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 also want to check if we can use those parameters to predict the other 2 datasets. Firstly we showed that we can accurately predict the other dataset of Finkelstein (the block mismatches) (Figure 15). We see again that the first part (till nucleotide 10) is predicted correctly and the second part shows more deviation from the data. This can again be explained by the fact that equilibrium is not reached after 10 minutes.

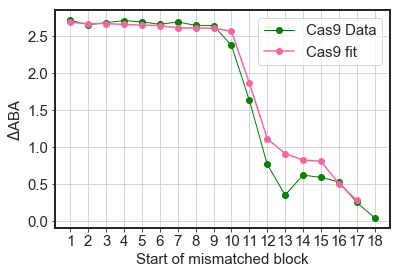


Figure : Prediction of the ABA for blockmismathces, based on the parameters obtained from fitting on single and double mismatches.

However, predicting the dataset of Boyle was less successful (Figure …). Qualitatively we can predict the pattern, since we see that the same patterns are conserved. But they are all way off from the actual values. …

The reason for that we cannot go from Finkelstein to Boyle comes again down to the assumption of equilibrium. If equilibrium is reached, all the occupancies should be exactly on the Hill-curve (except for some noise included in the measurement). However, if equilibrium is not reached, 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 we calculate the using the curve\_fit() function in python. This function uses the least squares method to find the best value for . But this means that there are multiple possible combinations of occupancies that will give the same . Therefore it becomes possible that our parameters do indeed correctly describe the value of the data, but at the same time describe completely different occupancies than the data does. So although it correctly predicts the , 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 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.

* Not being in equilibrium would also immediately explain why we could not predict the single mismatches with the parameters of the fit on Boyle.

Absolute ABA finkelstein is not the same as the absolute aba Boyle…

Maybe the concentration is not the concentration they thought they have

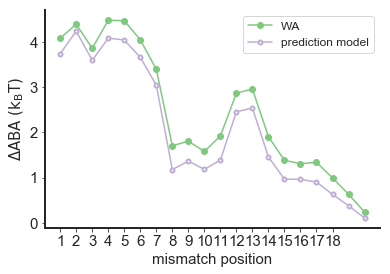
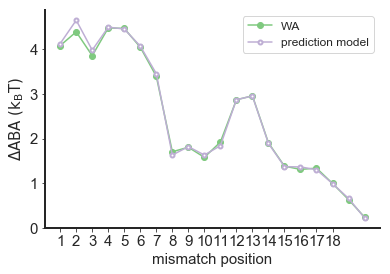
1. Derived properties of Cas12

Now we are able to use this kinetic model for Cas9, we hoped we could use this same algorithm for Cas12. Beforehand we knew that one adjustment had to be made, since we only have the data for Cas12 and not the absolute . This results in the fact that the on-target value is not known and this has to be fitted as well. However, the algorithm turned out to find multiple results with different on-target and values but comparable values. All the fits do give the same shape for the on-target energy landscape, it only is shifted by . This means that we would need the absolute ABA data to determine the height of the on-target energy landscape.

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. Unless we discovered that there are more complications than we expected on first hand, we are able to distract some features of Cas12, which will be described in 6.3 and 6.4.

# Fits based on 3 concentration points

For Cas 12 we do not have the intensity measurements for all the different concentrations, only the 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 plots for calculating the ABA with 3 or 8 concentration points, as can be seen in Figure 15. Therefore we now know that the concentrations differ at which the dynamics of different Cas proteins is measurable. So you do need to measure and fit on all concentrations if you move to a new type of Cas protein.

**A)** **B)**

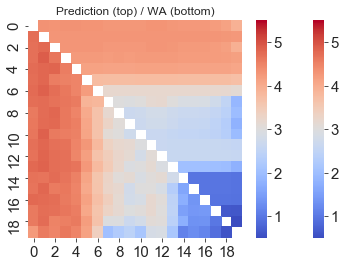
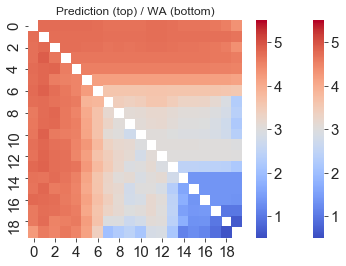
**C)** **D)**

Figure : A&C) with 8 concentration points. B&D) with 3 concentration points

# Assumption of equilibrium

For Cas12 we also want to check, if the assumption of reaching an equilibrium after 10 minutes, is met. Figure 16A shows that the correlation between the predicted values is almost perfect, except for the on-target. Also the predicted values for double mismatches do completely agree, even for the end of the sequences (Figure 16B). This indicates that on average equilibrium should be reached after 10 minutes.

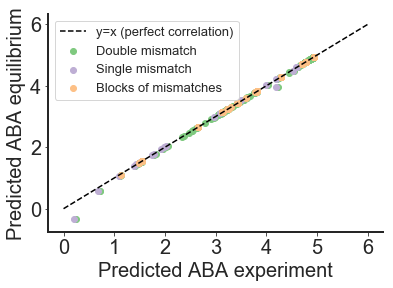
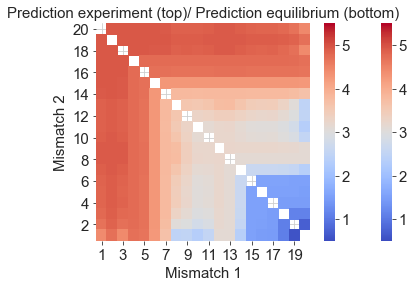
**A)****B)**

Figure : A) The correlation between the predicted values of the ABA based on the experimental calculation or assumed equilibrium. B) The prediction for the double mismatches using both calculations. The calculations are based on the parameters from the best fit of the fits against the weithed avarage of the data.

However, performing the fit using the assumption of equilibrium and calculating the according to equation (26), 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 17B). This second options also occurred for the fit where we follow the experimental calculations, only not that often (Figure 17A).

This second option found by the fitting algorithm has a much higher value (>100 000 instead of around 50 000) and does not show much similarity with the weighted average of the data. The parameters are fitted completely differently and the pattern in the parameters that we could distinguish before is completely vanished for this 2nd group of fits.

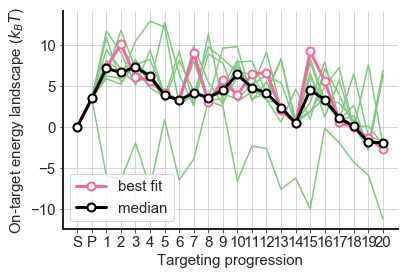
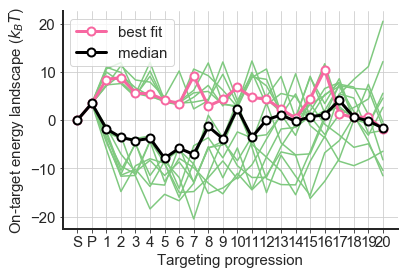
**A)** **B)**

Figure : The ontarget energylandschape. A) Calculations during the fit based on experimental procedure (3 concentration points) B) Calculations During the fit based on assumed equilibrium (3 concentration points)

The fits belonging to the 1st group do give good predictions for the single and double mismatches and the best fit has a value of 49011 which is comparable with the best fit without equilibrium ( = 49181). This leads to the conclusion that equilibrium is indeed met as we saw in Figure 16. 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 DO YOU SEE THIS HAPPENING MORE OFTEN IN EQUILIBRIUM???

# A closer look at the on-target energy landscape

In Figure 19 the on-target energy landscape for Cas12 is depicted for multiple fits, selected based on their value. It can be observed that they all agree at a few spositions: nucleotide 5,6 and 14. But besides that they have a lot of peaks at different locations. After manually adjusting the height and locations of those peaks and observing how the prediction power of the fit changes, we discovered that the location of the peaks does not have that much influence, as long as there are 2 bumps 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 this points is therefore not clear. For a possible explanation for the landscape being this undetermined, see section 6.2.2.

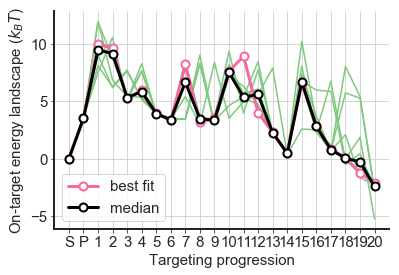


Figure : The on-target energy landscape for fits with a <3000 (compared to the weighted average)

# 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 20A.

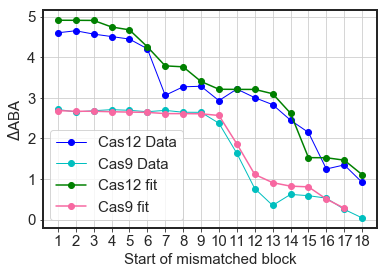
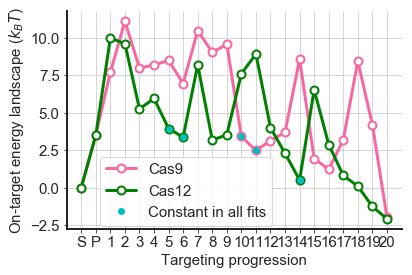
 

Figure : A) The predicted ABA for different block mismatches. B) The on-target energy landscape for Cas9 and Cas12.

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 to having it in another position in that region. Only from when the graph goes steeply down (starts at nucleotide 10/11), obtaining a mismatch would have less effect. Therefore, we can clearly define a so called ‘seed region’ for Cas9, going from nucleotide 1 till 10. 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. This also results in the fact that the on-target energy landscape is less defined, because since there is not a spot after which a mismatch matters much more or less, the position 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 has to 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 of all, leaving more characteristics of the on-target energy landscape undecided.

1. Discussion and outlook

….

Find right simulated annealing settings for Cas12

1. 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. However, it does not contain information to …

…

Since qualitative statements could be made, we did try to fit the algorithm to the data of Cas12. This led to the discovery that the dynamics of Cas12 gets observable at different concentrations than Cas9. Furthermore, the amount of time that is needed to reach equilibrium does differ for different Cas proteins. But even though equilibrium was reached for Cas12, adjustments in the settings of the algorithm are needed to make it properly work. In short, many more complications appear when you want to use the same model and algorithm for different Cas proteins.

Due to all these complications it was hard to define properties for Cas12. We did discover that Cas 12 does not have a clear seed region like Cas9 has. Therefore the location of the peaks in the on-target energy landscape does not have that much influence. Reasonable prediction power is obtained as long as there are 2 bumps visible in the landscape, with a valley in between of at least 3 nucleotides. Although it will 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.

1. 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 it is amazing research that they are doing and I am so glad I was able to take part in that.

For in the introduction:

conformational changes: ???? (other article Misha talked about?)

<https://www.annualreviews.org/doi/full/10.1146/annurev-biophys-062215-010822?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed>

* Loading guide RNA to cas9 complex (big)
* Upon binding to target DNA and PAM sequence (small)

Model, see site + figure 6 on site.

So if binding with target DNA and PAM sequence is only small, we do not take this into account. Our model starts from already loaded Cas9.

So only big difference with Zhang is the binding to PAM??

MATHTYPE

Difference of subtypes cas?

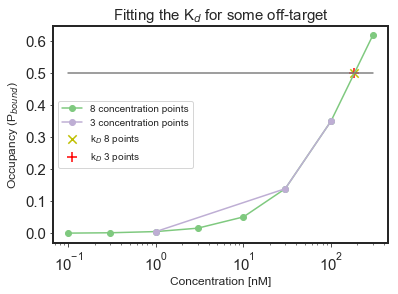
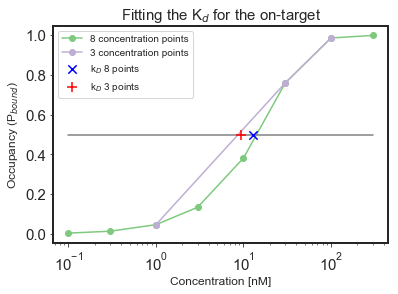
Assuming equilibrium 🡪 yes!

The 3 concentration points we use for the fit are not accurate to describe the data 🡪 check Cas9

It gives exactly the same plot for everything except for the ontarget-value. WHY????

NOT ONLY FOR THE ONTARGET! ALSO FOR OTHERS

You do not fit the ontarget during the fit, this is not in xdata.

1. Bibliography

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

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

[3] 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;

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

[5] 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;

[6] 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;

[7] 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;

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

[9] 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;

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

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

[12] 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;

[13] 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;

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

[15] 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;

[16] 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;

[17] 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;

[18] 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;

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

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

[21] 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;

[22] 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;

[23] 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;

[24] 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.

[25] 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

[26] 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;

[27] 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;

[28] 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;

[29] 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.

[30] 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

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

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

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

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