Table of Contents

[Introduction to CRISPR: 2](#_Toc10029726)

[The biology of CRISPR-Cas 2](#_Toc10029727)

[CRISPR-Cas as a gene-editing tool 3](#_Toc10029728)

[The focus of this thesis 5](#_Toc10029729)

[Explanation of the model 6](#_Toc10029730)

[The general concept 6](#_Toc10029731)

[The kinetics 6](#_Toc10029732)

[Enforcing a binding equilibrium 8](#_Toc10029733)

[Other assumptions 8](#_Toc10029734)

[Description of the different experimental set-ups 9](#_Toc10029735)

[Finkelstein et al.: 9](#_Toc10029736)

[Processing of the data from Finkelstein et al. 10](#_Toc10029737)

[Boyle et al.: 11](#_Toc10029738)

[Derivation of the Hill equation: ….. 12](#_Toc10029739)

[What we actually fit 13](#_Toc10029740)

[Simulated Annealing 13](#_Toc10029741)

[Selection of the fits 15](#_Toc10029742)

[Discovered properties of the data by fitting 16](#_Toc10029743)

[The usage of less concentration points 17](#_Toc10029744)

[Assuming equilibrium 18](#_Toc10029745)

[Comparison to Boyle 18](#_Toc10029746)

[Derived properties of Cas12 19](#_Toc10029747)

[Bibliography 21](#_Toc10029748)

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 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 an 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

It probably became clear that CRISPR-Cas is a widely used gene-editing tool. But how does it exactly work? CRISPR-Cas is originally a defend 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 which 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 the bacterium. This continues until the bacterium bursts open and releases al 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 later used as well, now in order to distinguish between viral DNA and the CRISPR-array that 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. It depends on the Cas protein how it cuts the DNA. Cas9 generates a double-stranded break 3 bp upstream of the PAM (13).

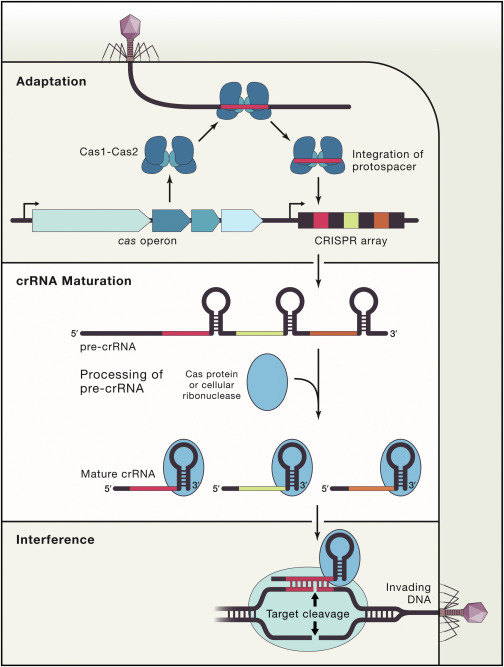


Figure : The three stages of CRISPR immunity (11)

We have seen that the CRISPR-Cas system exists of the CRISPR array (the memory of the system) and the Cas protein (the interference machinery). This Cas protein is an RNA-guided nuclease (RGN), which is a collective name of restriction enzymes that cleave the DNA only at sites that are complementary to the guide RNA. There are multiple of those Cas proteins (ex. Cas9, Cas12). Therefore the CRISPR-Cas systems have been assigned to 2 classes, which are further subdivided into 6 types and several subtypes (14). So far, CRISPR-Cas9 is the one that has been studied (and used) the most, which is a class II type II system.

# CRISPR-Cas as a gene-editing tool

A series of studies (15–19) led to the realization that induced DNA double-strand breaks (DSBs) can stimulate genome editing through homologous recombination events (20). This inducing of DSBs, is as we described above, 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 human embryos (1).

Therefore, we want to understand this off-target binding better. Studies from Fu et al. (21), Hsu et al. (22), Mali et al. (23), and Pattanayak et al. (24), 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. (25) and Finkelstein et al. (26) that we will use for this thesis. Hsu (22) and Pattanayak (24) also showed that mismatches appear to be better tolerated at high concentrations of Cas9. On top of that, Wu et al. (27) showed that Cas9 has indeed 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. (22) and a model from Stemmer et al. (28), use empirically determined scoring algorithms to quantify off-target cleavage. Doench et al. (29) proposed the cutting frequency determination (CFD) score to calculate the off-target potential of sgRNA-DNA interactions. Moreover, Singh et al. (30), build a model that solely uses DNA sequence information and trained their algorithm on experimentally validated data sets. Also more physical approaches exist. In a recently published paper, Zhang et al. (31) show their model based on the free-energy scheme of Cas9, of which they hope that it will be transferable to other Cas proteins as well, since it is based on solely physics. So far, no generally accepted model exists.

Building a model for CRISPR-Cas is what the group of Martin Depken at TU Delft is trying to do 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. (31), but we use a different ‘general model concept’ and are able to distract more features of Cas9. The advantage of having a physical 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 untill you get a ‘good’ model. This provides you with a lot of understanding about important features of the system, since you will discover which information will add predictive power to your model and what will not.
* A physical 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 (32). This model is trained on the experimental data of Boyle et al. (25) 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. (26) 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. (25). If this will give indeed 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 do expect that this model is 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:

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

We will first take a look at the differences between the two experiments from Boyle et al. and Finkelstein et al. in chapter … . Then we will take a look into the general concept of the model in chapter … . Thereafter we will take a look at the results from the fit based on the data from Finkelstein et al. …..

Explanation of the model

The off-target activity of CRISPR-Cas forms a big problem for using it as a tool to modify DNA. Misha Klein and Behrouz Eslami-Mossallam from the Martin Depken group at TU Delft started building a model, to gain mechanistic understanding of the targeting rules, by kinetically modelling the physics of guide-target hybrid formation (32). The main idea: View the hybrid formation between the guide RNA and target DNA as a random walk through a free-energy landscape. A match between the guide RNA and DNA will result in a lower free energy, a mismatch in a higher one. The thing that will happen (cleavage or unbinding) gets determined by the surrounding energy barriers. The difference in heights form the kinetic biases. They 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 the way we modelled this general concept. 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 the main goal is to understand the off-targeting activity of CRISPR-Cas9, we start the model only with what we already know and try to build from there, according to the laws of statistical physics. We know from experiments …(ref)… that Cas9 first needs to bind to the PAM region, then bind to all the other 20 nucleotides in the guide RNA, before it can finally cut. All those parts can be seen as 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 ). So what we get is a model with 23 states, where you can switch between the states (back and forward) with certain rates. To cut you have to go through all the states, until the last one where you cut, 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 how we modelled 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 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 level. The free energy is defined as: . E is the internal energy and S the conformational entropy at a fixed temperature T.

A close up of a map

Description automatically generated

Figure : … (32)

We use kinetic modelling, so we assume that for any single reaction to get completed is exponentially distributed. Therefore, the probability density for the reaction time of individual reactions 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 state. From this the chance to be in a specific state () can be derived:

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:

This also shows that the lower the energy barrier, the larger the rate, so the higher the change is to proceed in that direction.

Now we know how the rates between the states relate to the free energy landscape. But we can also say something about the chance to be in one of the states. The probability that a CRISPR-Cas system belongs to one of the states over time can vary, but we always know that it must be in one of the states (eq …).

Moreover, we can say something about how the probability changes for being in one of the 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.

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

This equation we call the Master equation. Solving it gives all the concentrations of reactants and products for any particular reaction pathway. For this model it is plotted in Figure 4 . 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 …. 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 occur during the whole experiment), we impose a detailed balance condition as stated in equation … . This is obtained by saying that the occupancy over time should not change if there is a binding equilibrium (see equation .. and ...).

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

# Other assumptions

The rate from solution to PAM is linearly dependent on the concentration of CRISPR-Cas9

All the internal forward rates are equal, the backward rate is determined by the relation as described above. So they depend on the energy landscape

……………………………..

Description of the different experimental set-ups

The physical model of CRISPR-Cas from Klein et al. (32) is trained on the experimental data of Boyle et al. (25). This data was used since by only looking at the data you could already see certain patterns, indicating that you can actually subtract some features and 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 are parameters and variables Klein et al. also use in their model. 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. (26) 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 this data as well.

# Finkelstein et al.:

Finkelstein et al. (26) developed 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 ). 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. They let the fluorescent labelled proteins flow through. Using high-throughput fluorescence imaging (done with 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 (26)

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 (0 nM, 0.1 nM, 0.3 nM, 1 nM, 3 nM, 10 nM, 30 nM, 100 nM, 300 nM).

The following assumption 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 yet to bind so many nucleotides.

## Processing of the data from Finkelstein et al.

The data we got from Finkelstein et al. is the fluorescence intensity. From this one can calculate the occupancy as described in Equation … . Here is …. And ­­ 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 Formula … . In general terms, the Hill equation describes the fraction of macromolecules saturated by ligand as a function of the ligand concentration. For the justification of the use of the Hill-equation in our model, see the derivation in section … .The then reflects the concentration at which . (see Figure 6)

From this 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 values indicate stronger binding.

The ABA gives the amount of free energy for each DNA strand. Since from the experiment we also know if these strands from a match or mismatch with the guide DNA in the CRISPR-Cas9, we can now determine the amount of free energy that a mismatch on every nucleotide will add to the energy landscape and thereby obtain the energy landscape of CRISPR-cas9.

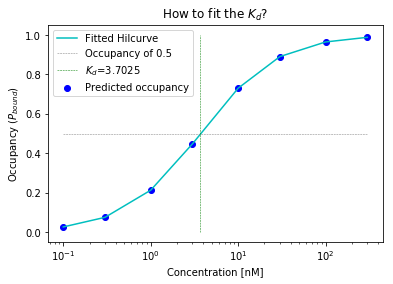


Figure : Based on the parameters the occupancy is predicted, through which the Hillcurve is fitted. at the concentration where the occupancy is equal to 0.5, the is found.

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). Since we enforce a binding equilibrium, we can say that the occupancy follows the Boltzmann distribution as described in equation … . Therefore the following holds:

()

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 …2…, 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:

Were we take:

* since our reference concentration in the experiments

Which leaves us with the Hill-equation:

WHERE EXACTLY DO YOU ASSUME EQUILIBRIUM???

# Boyle et al.:

Boyle et al. (25) 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. Only they measured the fluorescence for every strand at 3 different timepoints (500 s,1000 s, 1500 s) at a fixed concentration of 10 nM. The fluorescence intensity for the on-target was measured after 12 hours, to be sure that equilibrium was reached.

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.. 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 still in the linear domain and therefore reflect the on-rate. (Note that this effective on-rate that they measured is not the same as our rate from solution to PAM, it is the change in the occupancy.) To measure the off-rate, they 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, resulting in that fitting the straight line gives the off-rate.

In short, the datasets from the experiments from Boyle et al. (25) 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 justifies the use of linearization), 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.

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

In our fit we try to redo the experiment that Finkelstein has performed. We start by choosing a random set of parameters, with these parameters we can 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 … we can therefore calculate the backward rates from these and the randomly chosen forward rate. With both of these rates known we can build the rate matrix.

From the rate matrix we calculate the occupancy using the master equation (eq …). We do this for different concentrations. In doing so we have the exact same data as Finkelstein and we proceed in the same manner as described in section … to calculate the ABA. This can then be compared with the experimental data by taking the chi-squared value, 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 optimisation problem, this would be the fit with the lowest error (in our case the chi-squared).

One should only keep one thing in mind, if you cool down the system too quickly, it can get stuck in a local minima, instead of finding its global minima (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.

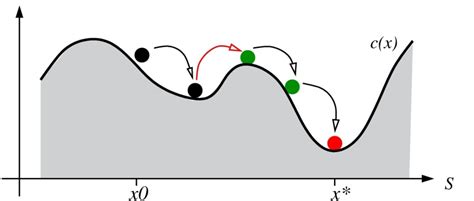
[](https://images.rapgenius.com/2706b9820eb743f64cb7da7226262e80.1000x426x1.jpg)

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 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 to quickly 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 use simulated annealing to achieve this. This whole process is done multiple times, since it is still possible to end up in a local minima instead of the global one. Therefore we need some kind of measure to tell which of the fits describes the data the best. The -value is one of those measures, which we do use during the fit to judge the quality of the fit. But our data is still a list with multiple measurements for the strand of DNA, so what do we consider the true value of the data here?

In biology the median is often used as the true value of the data, since you do not want to include outliers in your data analysis. However, when the data does not include those outliers and does not have a tail-like form, the mean can also be used. Since our data does indeed not include those properties…??, but we do have an measure error for every measurement, we will use the weighted average. To determine the weights that we need to use, we go back to the -function since this is the function that we are minimising during the fit. Therefore we say that the derivative of the -function should be equal to zero.

With these weights we can 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.

Discovered properties of the data by fitting

The first set of fits performed used all the data that was available from Finkelstein et al. (26). This means the for all double and single mismatches at (almost) every possible mismatch positions, for 8 concentration points (we excluded 0 nM since this would result in a division by 0), with the uncertainty in the measurement. The best fit seemed to get 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. This difference is caused by the amount of points we fit, for Boyle et al. (25) this were only 3 timepoints, while for Finkelstein we have 8 concentration points. Those 60 hours is way 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 us at the same time give us more insight in the data:

* *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:

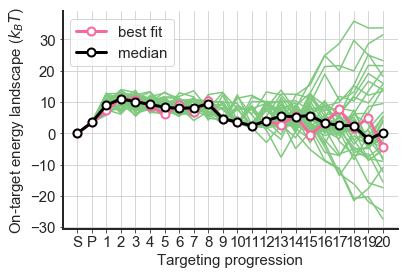


Figure : The first run of fits to the data of Finkelstein et al.. 50 fits are projected, showing a lot of diviation in the on-target Free energy landscape

# The usage of 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 than used in the Hill-equation to find the and calculate the as was explained in section … (processing the data). For our previous fit this was only done with all the 8 concentration points, but now we will test if we can get the same result if we use less concentration points. We will look at 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 once with the lowest error (see Figure 9A) which in this case remains 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 less concentration points is that it has the curve\_fit() function in python has more difficulties with fitting the Hill-curve through these points, resulting 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.

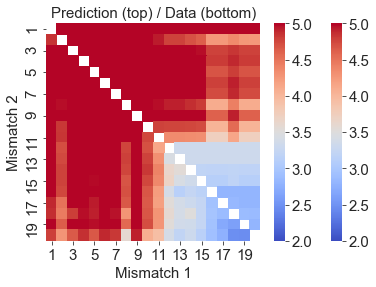
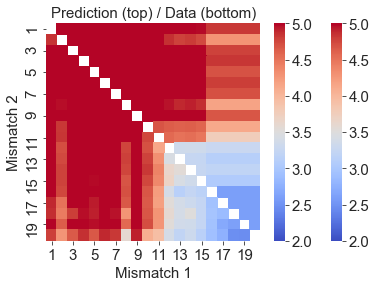
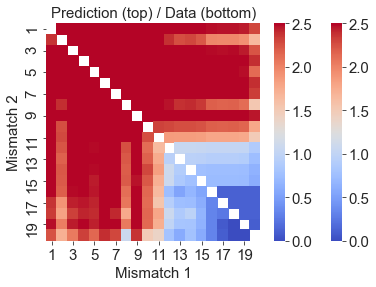


Figure : The double mismatch heatmap for: A) 9 concnetration points, b) 2 concentration points, c) 3 concentration points

# 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 …). Therefore, if equilibrium is actually reached, we can substitute this in the equation for ABA to obtain equation … .

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 anymore to do the calculation with the rate matrix (equation …) and the fitting to the Hill-curve (eq …), which will make the fit a lot faster.

CANT WE TAKE THE PARAMETERS AND CALCULATE ABA IN BOTH WAYS AND THEREBY SAY IF EQUILIBRIUM IS REACHED OR NOT???

By enforcing equilibrium in this way, we will immediately know if equilibrium is actually reached, by comparing this results.

The results form the fits….

…

Comparing the fit with and without equilibrium:

* What happens is that it just throws landscape up a lot. So it has higher rates, so it can still overcome it.
* Can equilibrium be assumed?

# Comparison to Boyle

….

2 state model idea behind the rates ???

Derived properties of Cas12

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

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. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR-Cas systems. Nature Reviews Microbiology. 2011.

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

16. 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;

17. 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;

18. 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;

19. 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;

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

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

22. 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;

23. 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;

24. 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;

25. 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.

26. 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]. 2017;170(1):35–47.e13. Available from: http://dx.doi.org/10.1016/j.cell.2017.05.044

27. 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;

28. 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;

29. 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;

30. Singh R, Kuscu C, Quinlan A, Qi Y, Adli M. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. Nucleic Acids Res. 2015;

31. 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.

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

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/