Right place, Right Time

Modeling the search time and specificity of Cas9 and Argonaute

Klein, Misha

DOI
10.4233/uuid:a27816fb-b042-4532-9304-4c3a675fd78f

Publication date
2019

Document Version
Final published version

Citation (APA)
Klein, M. (2019). *Right place, Right Time: Modeling the search time and specificity of Cas9 and Argonaute*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:a27816fb-b042-4532-9304-4c3a675fd78f

Right place, Right time

Modeling the search time and specificity of Cas9 and Argonaute

Right place, Right time

Modeling the search time and specificity of Cas9 and Argonaute

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,Prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Friday 13 December 2019 at 12:30 o'clock

by

Misha KLEIN

Master of science in Applied Physics
Delft University of Technology, The Netherlands
born in Chicago, USA.

This dissertation has been approved by the:

Promotor: dr. C. Joo
Copromotor: dr. S.M. Depken

Composition of the doctoral committee:

Rector Magnificus	chairperson
Dr. C. Joo	Delft University of Technology, promotor
Dr. S.M. Depken	Delft University of Technology, copromotor

Independent members:

Prof. dr. ir. S.J. Tans	Delft University of Technology
Prof. dr. I. Finkelstein	University of Texas at Austin
Prof. dr. D. Rueda	Imperial College
Prof. dr. P.R. ten Wolde	AMOLF/VU Amsterdam
Prof. dr. H. Schiessel	Leiden University

Reserve member :

Prof. dr. A.M. Dogterom	Delft University of Technology

Printed by: …

Front & Back: Sebastiaan de Ruiter

Casimir PhD Series, Delft-Leiden 2019-37

ISBN 978-90-8593-420-2

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

Contents

1

Introduction

Advances in genome engineering – 'making precise changes to DNA' – announced a new era of using Biology for Biotechnological applications. Most notably is the discovery of the CRISPR-Cas system which over the course of the past decade has facilitated the development of strategies for making drought-resistant crops, targeted antimicrobials, organ transfers from pigs to humans, eradicating malaria mosquitoes and more. In spite of CRISPR-Cas systems having become a common tool in many scientific laboratories, their application – especially treating humans – remains in its infancy due to concerns regarding its precision.

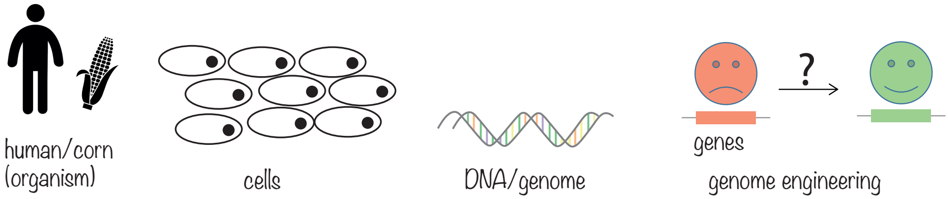


Figure 1.1: Living organisms inherit traits encoded in their DNA. Making (precise) changes to an organisms genome is called ’genome engineering’.

1.1. Genes, Genomes and Genetic Engineering

What does making a drought resistant crop have in common with treating sickle cell disease? To answer this question we must first consider what plants, humans, bacteria and all other living organisms on Earth have in common. All living organisms consist of cells (**Figure ??**). Inside the cell resides its genetic material: DNA. DNA is said to encode for ’genes’, resulting in an organism’s traits such as the color of an apple, or the color of our eyes. Other traits, such as the corn being drought resistant, or a human having a hemoglobin mutant leading to sickle cell disease, are also a direct result of the precise ’gene code’ or ’genome’. If we could somehow edit (’engineer’) an organism’s genome, we should thereby be able to change its traits. In case of our two examples, both traits (drought resistance in corn, human sickle cell disease) are caused by a single gene, and are thereby altered just by editing the associated gene. However, what if we by mistake edit the wrong gene? This could potentially have dire consequences.

In this thesis we are concerned with understanding the most novel genome engineering tools by means of mathematical and physical modeling. To understand how we go about translating editing specificity into physical quantities (time, energy, etc.) we must first take a deeper dive into their Biological origins.

1.2. Beyond the Central Dogma

The cell is the building block of all living tissue. Inside each cell countless of chemical reactions take place to make it grow, protect it against an ever changing environment, and eventually make it divide – giving rise to new life. Virtually all cellular processes are carried out by molecules called proteins. Orchestrating all chemical reactions requires that the correct amount of active protein is available at the right time.

One way of achieving this goal is to control the amount of each protein produced in the first place. The cell encodes these instructions in the form of another kind of molecule: Deoxyribonucleic acid or DNA for short. Encoding such information is possible since there are just four forms each monomer constituting one unit of either of the two helical DNA chains (strands) can take on: Adenosine (A), Thymine (T), Guanine (G) and Cytosine (C), called nucleotides. Moreover, in forming the double-stranded DNA (dsDNA) of an organism’s genome every A-nucleotide positions itself opposite to a T-nucleotide. Similarly, a ‘C’ is said ’to form a base pair with a G’ (**Figure ??**). This base pairing property allows the

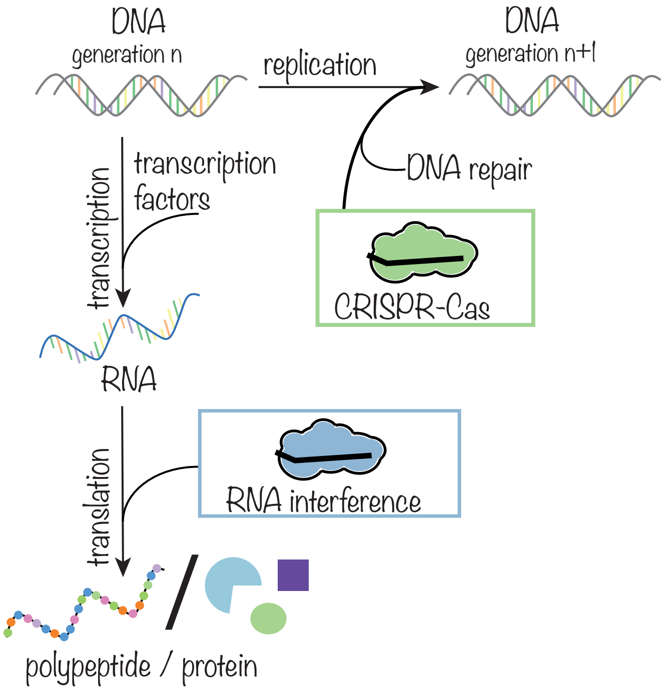


Figure 1.2: The Central Dogma of Molecular Biology states that genetic information is stored as DNA, copied during replication (cell division) and read out by first transcribing it into RNA and then translating it into an amino acid sequence. Each of these processes is heavily regulated by the cell. Target searching proteins play important parts herein: Transcription Factors act on transcription, DNA repair mechanisms safeguard replication. Non-coding RNA guided nucleases fall into this category as well: CRISPR associated (Cas) proteins prevent replication of viral elements, while RNAi controls translation levels.

cell to encode information in the DNA's nucleotide sequence, much like a computer stores information in binary sequences.

Processing of the encoded information, resulting in protein synthesis, happens in a series of Chemical pathways famously termed 'the Central Dogma of Biology' (**Figure ??**). During cell division each daughter cell acquires an identical copy of the mother cell's genome (DNA). As each cell only has a single copy of the genome, before cell division takes place the DNA gets replicated. To synthesize a protein, the DNA first gets transcribed, resulting in a precursor molecule, RNA, in which every nucleotide of one of the DNA's strands is replaced by its complement – with the exception of Thymine that gets replaced by Uracil (A to T, T to U, C to G and G to C) (**Figure ??**). These RNA molecules are now ready to get translated into a sequence of what are called amino acids that eventually folds into its final form: a protein. Amazingly, all steps within the Central Dogma – replication, transcription and translation – are actually carried out by proteins themselves.

Although the Central Dogma in essence details the flow of genetic information from DNA

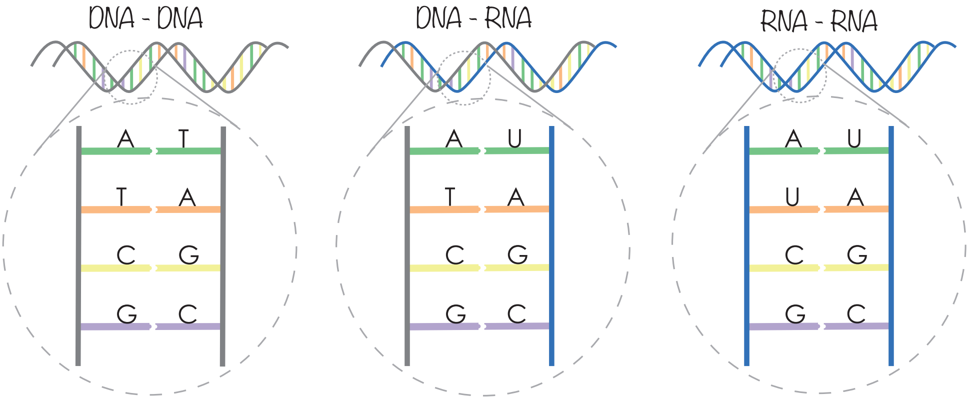


Figure 1.3: Double-stranded nucleic acids form ’base pairs’. (left) DNA-DNA pairs, A(denine) complements T(hymine) and C(ytosine) complements G(uanine). (middle) DNA-RNA pairs, U(racil) replaces T(hymine), thereby matching A(denine). The DNA’s Thymine still matches the RNA’s Adenine. (right) RNA-RNA pairs, A(denine) complements U(racil). We will refer to any of the shown base pairs as ’matches’, while any other possible pair (i.e. A-G) as ’mismatches’.

to protein, more detailed control of protein levels is achieved through numerous ‘feedback loops’. When DNA damage occurs, a set of proteins involved in DNA repair mechanisms must recognize and restore the original sequence to avoid transcribing incorrect instructions or passing them on during replication. Transcription levels (the amount of RNA produced by a particular gene) are actively up- or down-regulated by proteins termed ‘transcription factors’ that bind near the gene of interest to either facilitate or repress the proteins that carry out transcription.
In this thesis we shall focus on a different kind of regulation that uses so called non-coding RNA molecules. Unlike originally envisioned, RNA molecules are more than merely intermediates between DNA and protein. In fact, large portions of the genome do not even directly encode for proteins at all – estimated to be more than 95% of the human DNA [**?**]. Partially, DNA can encode for RNA that is not meant to be translated: ‘non-coding RNA’. Instead, making these RNA molecules bind to specific RNA or DNA sequences, using the base pairing rules mentioned above, can direct proteins to catalyze reactions at desired sequences only. Examples can be found throughout all major kingdoms of life. Eukaryotes – amongst which yeast, plants, animals and humans – make non-coding RNA bind to messenger RNA (mRNA, the coding form of RNA), thereby inhibiting its translation. Prokaryotes – archaea and bacteria – use these non-coding RNAs to detect invading viral DNA and signal it for destruction. We shall detail both below.

Taken together, the cell uses DNA to store and read out information. In turn, parts of the DNA are used to safeguard and regulate the flow of genetic information in an attempt to prevent mistakes during read out and protect the integrity of the host’ instructions. Recognition of specific DNA/RNA sequences plays a crucial role herein.
The remainder of this chapter briefly reviews some of the different classes of small non-coding RNA molecules, highlights their technological potential and explains the need for

physical modelling of the kind used throughout this thesis. The latter parts of this chapter contain a brief introduction into the relevant theoretical techniques that allow us to couple the physics to experimental data. Although not needed to understand '*the why*' and '*the what*', that section serves to additionally explain '*the how*' of all subsequent chapters.

1.3. Nucleic acid guided, nucleic acid effector complexes

1.3.1. The CRISPR-Cas adaptive immune system

Organisms have evolved various strategies to cope with their ever changing environments. This too holds true for even the smallest of organisms: Prokaryotes. For bacteria and archaea such threats are 'mobile genetic elements' (MGEs), foreign DNA (or RNA) originating from either viruses or plasmids. Bacteriophages, viruses that invade bacteria, in essence consist of no more than a container with their genetic material. They do not possess the required protein machinery to read out their own DNA. Hence, by themselves, they are incapable of replicating. For this reason, phages 'invade' host bacteria by injecting their DNA into them, hoping that the bacteria will not recognize it as being foreign and proceed to transcribe and translate it as if it being part of its own genome. The viral genome will encode for the proteins of the DNA-containing capsids that make up the body of the phage particle. If too many of such new phage particles get synthesized inside the host, the internal pressure can increase to such levels that the bacteria will burst open, setting the new virus particles free, enabling them to invade new bacteria.

Despite bacteriophages being about ten times more abundant, their prokaryotic hosts are still one of the most abundant life forms on earth [**? ? ?**]. Prokaryotes have, akin to what we know from humans, evolved immune systems. The centerpiece of this thesis is an adaptive immune system (meaning it adjusts to the incoming phage as opposed to innate systems that use a generic defense response) discovered in about half of all sequenced bacteria species and nearly 90% of all archaea [**? ?**]. About a decade before its function became clear, researchers discovered a particular set of non-coding sequences as part of the bacterial genome. The bacteria encode for an array of partially palindromic, more conserved, sequences. These 'repeats' are separated by highly variable sequences ('spacers'). It took until the early 2000's to realize the origin of these spacer sequences. Pioneering bioinformatics research found the spacer sequences to be originating predominantly from MGEs [**?**]. Soon after followed the first experiments demonstrating how this is part of an adaptive immune response [**?**]. The authors challenged phage sensitive *S. thermophilis* bacteria to new phages. Remarkably, the bacteria were able to survive the new attack. In addition, bacteria that became immune did indeed incorporate a novel spacer sequence from the phage into their, as it is now known to be, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) array. Later experiments revealed the roles of a set of proteins, typically co-transcribed with the array, termed CRISPR associated (Cas) in acquiring the new spacers, processing those spacers into guides and the destruction of the phage's DNA (**Figure ??**).

Upon encounter of a new phage genome, a set of Cas proteins – Cas1 and Cas2 – acquire the new spacer and incorporate it into the array [**?**](step 1 in **Figure ??**). Together, the Cas1-Cas2 complex (at least the variant found in *E.coli.*) adapts a butterfly-like structure that neatly fits a single spacer sequence. Next, the CRISPR locus gets transcribed, the cas genes get translated, while the CRISPR array forms non-coding RNA. Processing of the array's transcript results in small RNA fragments that contain the transcript of individual spacers [**?**](step 2). These small RNAs get loaded into either a single or a complex of Cas protein(s) [**?** **?**](step 3). Note that after transcription the resulting RNA actually contains the sequence complementary to the DNA it originates from (see base pairing rules, **Figure ??**). Therefore, this 'guide RNA' (gRNA) is able to direct the Cas protein to the viral DNA site. Once bound, the loaded Cas protein either possesses or recruits a nuclease (a DNA cleaving enzyme) to destroy the viral DNA (step 4 in **Figure ??**) [**?** **?** **?** **?**].

For the CRISPR system to convey immunity to its host it must do more than effectively degrade or inactivate the foreign DNA. It must be able to distinguish self- from non-self (DNA in this case), preventing self-targeting, otherwise called autoimmunity. Partially this requirement is met by using the spacer sequence to generate the guide RNA. However, by construction the CRISPR array itself contains a perfect copy of the target. Also, the host' DNA, by chance, may still contain a sequence similar to the spacer outside of its CRISPR array. If the bacteria was to target its own DNA, it could kill itself. Most CRISPR systems prevent this by pre-selecting spacers that are preceded by a short (typically 3-5nt) motif termed the protospacer adjacent motif (PAM) [**?** **?** **?**]. Only the protospacer (the sequence complementary to the spacer on the opposite viral DNA strand) and not the repeat sequence in the CRISPR array is flanked by the PAM. Direct interactions between the Cas protein and DNA can determine whether the DNA is foreign and should be marked for destruction. A wide diversity of CRISPR-Cas systems have been discovered thus far. Despite the zoo of different loci (sub-)types – 19 subtypes and still counting – they share a common architecture (**Figure ??**). The CRISPR array, the memory of past infections, is co-transcribed with the cas genes. As mentioned, integration of the novel spacers, adaptation, into the array requires the proteins Cas1 and Cas2.

To classify the different CRISPR systems a two-step scheme is currently used [**?**] (**Figure ??**). The first step groups the CRISPR loci into one of two major classes. Class I systems use a multi-subunit protein complex for targeting and degradation of the foreign DNA (interference), whereas in class II systems this is carried out by a single Cas protein. The second layer of classification is based on the presence of signature Cas proteins. Class I type I systems, the most abundant subtype, use a mixture of the proteins Cas5 through Cas8 to form a larger protein complex termed Cascade ("CRISPR associated complex for anti-viral defense") [**?**] that uses the crRNA guide to bind to the viral DNA. Once bound, it recruits yet another protein: Cas3, the signature protein of type I systems, that is able to unwind and degrade the phage genome [**?**]. Similarly, type III systems form an interference complex from the proteins Cas5 through Cas7 and their signature protein Cas10.

Class II systems (types II, V and VI) are considerably less complicated. Target interference is carried out by a single Cas protein (see Cas9,12-14 in **Figure ??**) that possesses nuclease ('cleaving') domains. For this reason, class II CRISPR systems are particularly interesting from a technological perspective, as shall be highlighted below.

Even amongst CRISPR systems of the same type (and therefore class), there exist significant differences. Such subtypes can vary based on differences in size or function of their signature gene or contain additional non-signature Cas genes, a prime example being Cas4 which recently has been found to take part in the adaptation process for type I-F systems [**?**

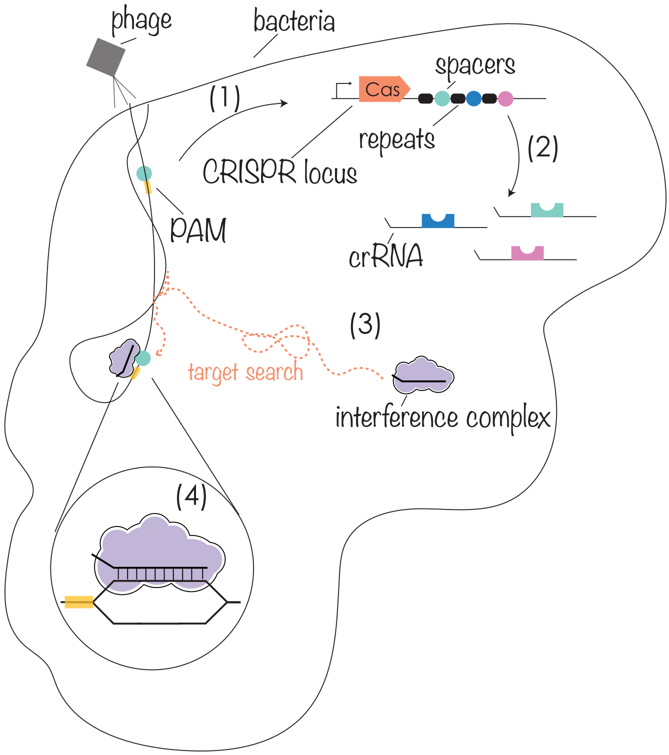


Figure 1.4: The CRISPR-Cas system provides immunity against invading bacteriophages. (1) Upon infection novel spacers are aquired from phage DNA and incorporated in the host' CRISPR locus. (2) Transcription and further biogenisis results in CRISPR-RNA (crRNA) guides. (3) Cas nucleases loaded with the crRNA can search the invading genome for matches to the guide (colored dot) that lie adjacent to a PAM sequence (yellow rectangle). (4) Having found a proper target, the Cas nuclease binds the DNA stabily and becomes cleavage competent.

]. Moreover, new CRISPR systems are still being discovered, such as the subtypes of type V that use the protein Cas14 [**?**].

In a nutshell, the CRISPR-Cas system uses RNA guided Cas proteins to perform sequence specific DNA edits.

1.3.2. RNA interference

Gene regulation, tuning the amount of protein produced from a given gene, is essential to any living organism. Cells partially achieve this by controlling the transcription levels of every gene. Additionally, post-transcriptional regulation is in place that modulates translation levels. Over 60% of all the protein encoding mRNA in human cells is subjected to a type of regulation known as RNA interference (RNAi) [**?**]. Eukaryotic systems possess several RNAi pathways characterized by the form of the non-coding RNA it utilizes (**Figure ??**) [**? ? ? ?**].

Mammalian genomes partially encode for non-coding RNA termed pri-microRNA (step 1 in

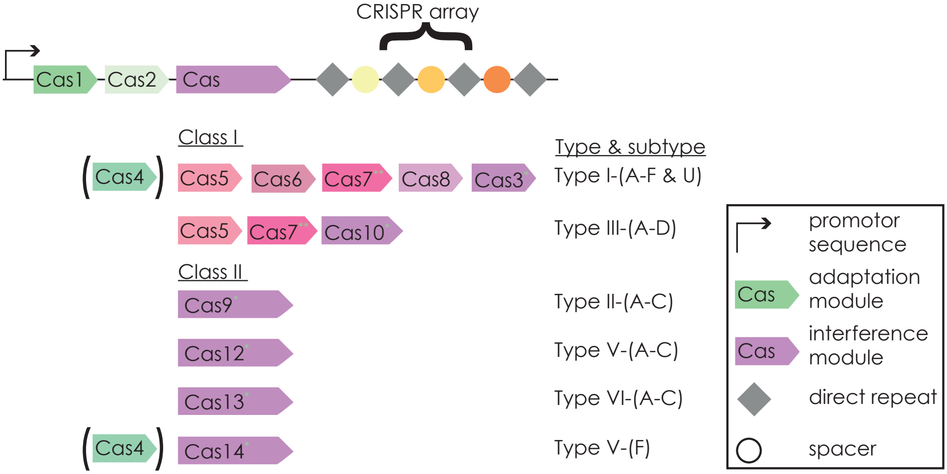


Figure 1.5: Classification of CRISPR systems. Typical architecture of the CRISPR locus is shown on top: The operon controls the adaptation and interference machinery as well as the CRISPR array. Below the most important differences between different types of CRISPR systems in their adaptation/interference modules are shown. For a more elaborate list of CRISPR (sub-)types see [**?**]. * signature gene, ** multiple copies present on locus.

Figure ??A). These long transcripts are processed inside the nucleus by a protein named Drosha, resulting in pre-microRNA (step 2). Exporting the pre-microRNA outside the nucleus, into the cytosol, and further processing by the protein Dicer produces the final microRNA that contains the information needed to silence the translation of a mRNA (step 3). The microRNA guide molecule gets loaded into a protein termed Argonaute (Ago), forming a RISC ("RNA-induced silencing complex") (step 4). As discussed above, the CRISPR system uses the crRNA to guide Cas molecules to their complementary target. Similarly, a microRNA-loaded Ago protein binds to mRNA at what is termed the 3' untranslated region (3'-UTR), which as its name suggests serves as a demarcation of the stopping site of translation (step 5). By occupying the 3'-UTR, Ago blocks the translation machinery either directly or by recruiting co-factors that actively degrade the mRNA.
A second RNAi pathway produces small interfering RNA (siRNA) guides from double-stranded RNA (step 1 in **Figure ??**B). Such dsRNA, originating either from within the cell itself or from viral elements, reside in cytosol. The siRNA molecules are produced by Dicer (step 1) and loaded into Argonaute (step 2). The siRNA pathway can either function to inhibit transcription, the same way microRNAs are used, or target viral RNA (step 3).

Within the first few years after its initial discovery in 1998 [**?**], RNAi based therapeutics started to emerge in which either the siRNA or microRNA pathway is manually activated by injecting synthetically designed dsRNA into the cell to target specific mRNAs of interest. For this reason, its authors, Andrew Z. Fire and Craig C. Mello, received the 2006 Nobel prize in Physiology and Medicine [**?**] less than a decade after their original publication detailing this programmable RNA targeting system.

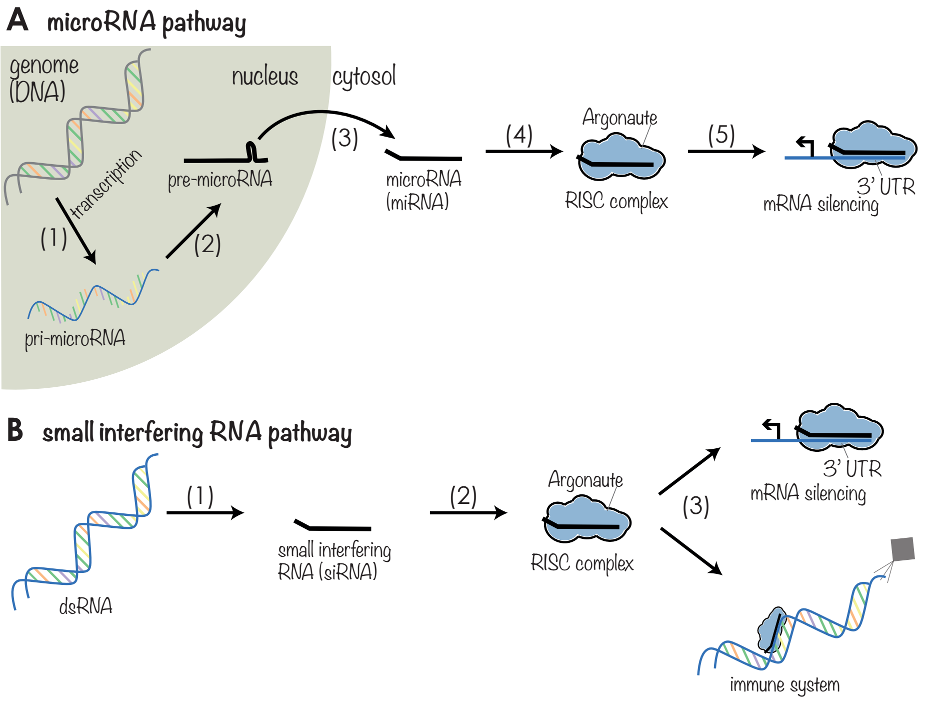


Figure 1.6: RNA interference pathways (in eukaryotes). **(A)** microRNA pathway:(1) non coding RNA encoded on the genome. (2) 'cropping' by Drosha. (3) Exporting by Exportin 5 and 'dicing' by Dicer. (4) Loading of the guide into Argonaute. (5) RISC complex silences messenger RNA by binding to the 3'-UTR. **(B)** siRNA pathway:(1) dsRNA in cytosol is processed by Dicer into siRNA.(2) Loading of siRNA into Argonaute. (3) RISC can either silence mRNA or fight-off invading (RNA) virusses.

Peculiarly, Argonaut proteins have also been found in prokaryotes. Due to their similarity to their eukaryotic counterparts, and the CRISPR systems described previously, these Ago proteins are speculated to be involved in gene regulation or anti-viral defense. However, in many such cases, their precise function remains elusive [**?**]. Regardless, after variants have being reported that use DNA guides and/or target DNA [**? ?**], researchers have been interested in exploring also Ago's potential for genome engineering applications.

1.4. The genome engineering toolbox

What if we could express Cas9 outside of its bacterial host and load it with a guide sequence we designed ourselves? Could we thereby target a DNA location of our choice? Researchers in 2012 have demonstrated exactly this. Type II CRISPR systems express a two-part RNA, consisting of what are termed the CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA). Jinek et al. [**?**] demonstrated that it is indeed possible to perform edits *in vitro* using a single synthetically designed guide RNA (single guide RNA or sgRNA). Soon after followed the first demonstration of editing human and mouse genomes [**? ?**]. These studies further utilized that Cas9 also preprocesses its guide from the CRISPR array's transcript (performing step 3 in **Figure ??**) [**?**]. Designing a DNA containing several guides, separated by repeats to form a ’synthetic CRISPR array’, the researchers demonstrated the ability to edit the (human) genome at multiple sites at once [**?**].

It is relatively inexpensive and simple to design a DNA guide to target a desired (DNA) target. Cas9-sgRNA systems readily became commercially available. It therefore did not take long before researchers would demonstrate CRISPR-Cas9 based genome editing can be done in virtually any organism of interest, ranging from typical model systems for Biological experiments as Drosophila (the fruit fly) to technologically relevant *E.coli*, crops and plants, livestock and, as mentioned, even human cells. CRISPR-Cas9 has shown the potential to be applied in numerous applications of which generating drought resistant plants [**?**], targeting antibiotic resistant bacteria [**?**] and treating genetic disorders [**?**] are just a few.

In essence, gene-editing uses Cas9 to cut an unwanted gene and relies on the DNA repair machinery to either simply ‘remove’ it or replace it with a sequence supplied externally (**Figure ??**). Other than CRISPR-Cas9, the ‘genome engineering toolbox’ is rapidly expanding with other guided DNA nucleases. For instance, Cas12 [**?**] and even some bacterial Ago [**?**] also enable DNA editing. Alternatively, nuclease inactive, or ‘dead’ dCas9 still binds DNA, but is engineered to not cut it. Fusing dCas9 to other (bio-)molecules can direct these to the desired sequence. For instance, fusing dCas9 to transcription factors can direct them to a gene of interest to ‘interfere’ or ‘activate’ them (CRISPRi/CRISPRa)[**?**], tuning transcription much like RNAi tunes translation (**Figure ??**). Instead, attaching fluorescent proteins to dCas9 allows one to illuminate a specific part of DNA [**?**] (**Figure ??**). It is even possible to tie the binding or cleavage by (d)Cas9, or the increasingly popular variant Cas13, to a visible change of the solution’s color [**? ? ?**] (**Figure ??**). These techniques allow one to detect small amounts of DNA from infectious diseases or genetic disorders.

1.5. Off-targeting

Unfortunately, RNA guided nucleases (RGNs) are not 100% specific. There are numerous studies demonstrating CRISPR-Cas9 [**? ?**] either binding or cutting target sequences that do not fully match their guide RNA (DNA-RNA pairs other than those shown in **Figure ??**). Given their Biological roles in immune systems, it is actually not that surprising. Viruses typically mutate extremely fast, meaning that any spacer sequence acquired by the CRISPR system would rapidly be outdated if it were not to also target slight variations of the spacer sequence. Moreover, bacterial genomes are about 1000 times shorter than mammalian genomes, increasing the probability of encountering off-target sites when repurposing CRISPR-Cas9 for human cells.

Unintentionally cutting DNA at an unwanted location can cause serious damage to the cell. In an attempt to counteract off-target activity as much as possible, different strategies have been demonstrated to work. For instance, one may search for a Cas9 other than that from the most common host (*streptococcus pyogenes* (spCas9)), or another CRISPR system altogether, such as Cas12, that naturally appears to exhibit less off-target activity [**? ?**]. Other strategies [**?**] include mutating Cas9 to make it light-inducible to limit its dosage , turning

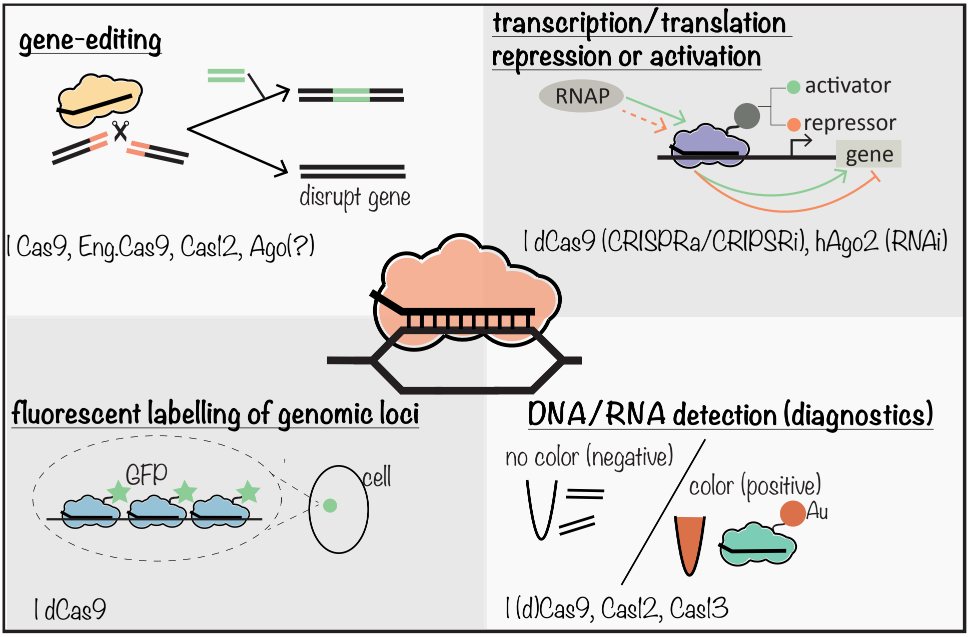

gene-editing
disrupt gene
| Cas9, Eng.Cas9, Cas12, Ago(?)
transcription/translation repression or activation
RNAP
activator
repressor
gene
| dCas9 (CRISPRa/CRIPSRi), hAgo2 (RNAi)
fluorescent labelling of genomic loci
GFP
cell
| dCas9
DNA/RNA detection (diagnostics)
no color (negative)
color (positive)
Au
| (d)Cas9, Cas12, Cas13

Figure 1.7: The "genome-engineering toolbox". CRISPR-Cas9, CRISPR-Cas12, CRISPR-Cas13, and even Argonaute are utilized in many different ways (see text). From the perspective of our model, the different systems are fairly similar: a protein loaded with a guide that targets the complementary sequence.

it into a nuclease only for single-stranded DNA (a 'nickase') or reducing the length of the guide RNA [**?**]. Using protein engineering even synthetically designed high-specificity Cas9 variants have been made [**? ? ?**].

The strategies above have proven to be successful. However, the major challenge reducing off-target activity faces is actually the detection of off-target activity itself. There is an immense amount of experiments needed to determine all off-targets for all possible guides, even for a single gene target (**Figure ??**). On top of that, detecting genome-wide off-targets for even a single Cas9-sgRNA has proven to be challenging. State-of-the art detection of genome-wide off-targeting unfortunately suffers from a rather low resolution [**? ?**]. The sequencing techniques used offer a detection limit around 0.1% - meaning 1 in a 1000 sequenced DNA must contain a cut. Note that this is still quite high compared to the shear amount of DNA present in all the cells of an organ(ism) combined. To further advance the application of CRISPR-Cas9 based gene editing, it is increasingly important to accurately tell more than these 'highly probable' events. Although cutting any particular off-target might happen infrequently, combining the possible billions of those events that may occur on a genome makes that some off-targeting is actually highly probable (**Figure ??**). Moreover, infrequent off-target events can be enough to cause serious damage or even disease.

Stepping away from genomic target sites, one can design an *in vitro* experiment that subjects Cas9-sgRNA to a library of off-targets containing a variety of mismatch patterns. Recent experimental techniques (data used in following chapters) use this to allow for accurately detecting the full range of activity. The hope is the outcome of such experiments can be translated back to the setting of an application, thereby avoiding the need to repeat these experiments for every possible guide of interest. To this end, several computer models are build based on the available data, with the goal of predicting off-targets.

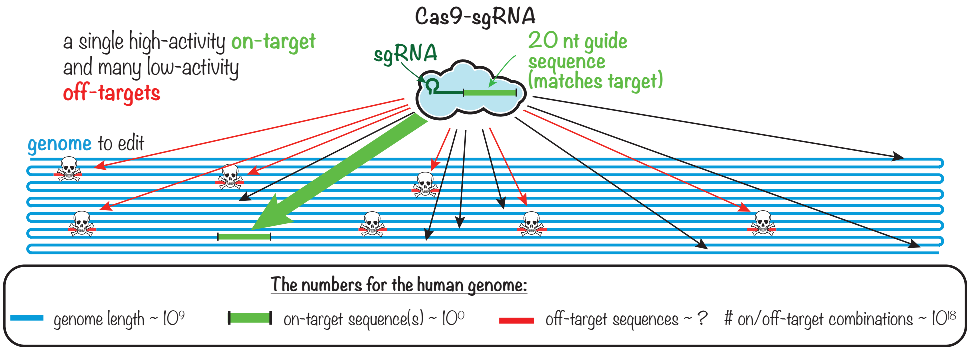


Figure 1.8: Although Cas9-sgRNA (or any other RGN of choice) predominantly targets the site that matches its guide (green), it is not perfect. The shear volume of low frequency off-targeting events makes encountering an off-target more probable. Although not all off-target edits are necessarily harmful (red, as opposed to the black arrows), a single mistake can have consequences to the cell. Numbers in the inset are upper back-of-the-envelope estimates assuming a random genome of human length, and are meant to demonstrate the imbalance between on-target and the vast number of off-targets.

1.5.1. Off-target prediction tools

A guide RNA sequence is only 20nt long. As a result, there are likely multiple different guide sequences that can be used to target a specific gene (typically thousands of kilobases). There exist several computer algorithms to decide which guide should be used to disrupt a particular gene locus (**Figure ??**). In essence, the user supplies a candidate guide sequence, the target sequence and the genome to be edited. The computer algorithm will return a list of (the most highly probable) off-targets. Their workings can be characterized into one of three types (see **Figure ??**). Alignment based prediction tools, such as CasOFF-finder [**?**], ChopChop [**?**] and E-CRISP [**?**], do no more than search for sequences on the genome that share sequence similarity to the intended target. Other tools use a mathematical model to score/rank the propensities for off-targets to be cut. The model incorporates empirically determined scoring schemes in a somewhat ad-hoc fashion. Examples include MIT's prediction tool [**?**], CCtop [**?**] and the Cutting Frequency Determination (CFD) score [**?**]. A third, and increasingly popular, category of prediction tools is based on Machine Learning [**? ?**] in which a large amount of data is used to build an AI-based decision tool.

Unfortunately, each of the mentioned prediction tools lacks good performance trying to predict experimentally determined genomic off-targets [**?**]. For this reason, it is becoming increasingly important to go beyond such 'data driven prediction' and better understand the processes by which RGNs search for and recognize their target site within a genome.

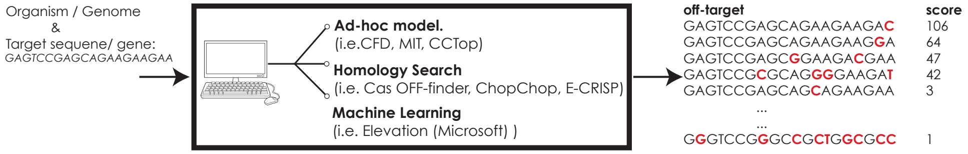


Figure 1.9: Existing predictions tools allow the user to provide the target gene locus & organism and output a ranked list of off-targets. Red nucleotides indicate mismatches.

1.6. A physics-based approach

Say, you want to edit a specific human gene. How likely are you to encounter an off-target, given a particular nuclease and guide? Or, say you want to build a diagnostics tool based on CRISPR-Cas9 (**Figure ??**). What is the expected false positive/negative rate of your design? To answer such questions, we must move beyond the aforementioned scoring schemes and build a quantitative model. Instead of only asking *if* a particular sequence will (likely) get cleaved, we additionally seek to understand *why* certain sequences are preferred - something none of the aforementioned prediction tools is capable of doing. More precisely, we ask:
”What fraction of DNA molecules with sequence X (typically) gets cut (or merely bound) if I subject my sample to a given concentration of Cas9-sgRNA for a specified time?” With such information, it becomes possible to use the computer to mimic any technique in which the RGN is applied to predict its read-out.
To do such we build a physics based model. Restricting our model to be governed by the laws of physics, as we would believe any experimental data to be, should in principle guarantee an accurate performance for both probable and infrequent (off-)targets. This should not only allow us to train our model using the existing data with highest signal-to-noise ratio, it should in principle require far less data all together. As shall become clear in later chapters, this allowed us to use datasets of lesser size, but higher quality, as our training set. There are several other benefits for using a physical model.
If we are able to pinpoint the correct physical laws governing the target interference, we should also be able to explain directly what feature in some sequence X makes it susceptible to cleavage, that some other sequence Y is lacking.
As building such a model necessitates a level of abstracting RGN systems (**Figure ??**), we will hopefully learn along the way precisely what targeting principles are shared . At the very least, fairly comparing RGNs (i.e. Cas9 and Cas12) will detail exactly what should be ‘the tool of choice’ for a particular situation (**Figure ??**).

Unfortunately, constructing a physical model of the target recognition process for a RGN is the hard part. What are the physical laws that are most important to incorporate and how to translate those into a mathematical model? The remainder of this thesis presents our best attempts at answering those questions.

1.7. Basics of physical modeling techniques

This section presents an overview of the physical theories and concepts used throughout this thesis. This is not meant as a necessary prerequisite for following any reasoning detailed in subsequent chapters, nor will it be needed to understand any conclusions thereof. Instead, the collection of topics discussed here form the basis of all mathematical derivations and simulation techniques used.

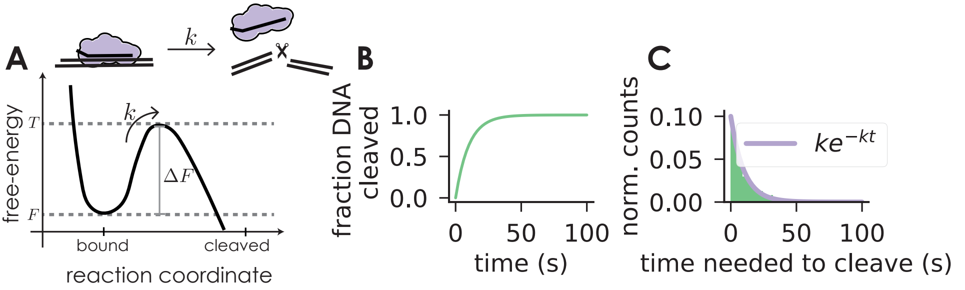


Figure 1.10: A single chemical reaction in which the RGN cleaves its substrate at rate k. **(A)** Free-energy landscape. A barrier of height ΔF - the distance from the bound state's free-energy (F) to the least favourable intermediate (the transition state T) - separates the bound from cleaved configurations. **(B)** Population of cleaved DNA as a function of time. **(C)** Reaction time for individual reactions - histogram produced generating many realisations - are exponentially distributed.

1.7.1. Kinetics 101

The cell can be viewed as a busy chemical factory. Molecules move around, occasionally colliding into one another, enabling them to exchange chemical bonds, leading to new chemical species. Technically, any such reaction is thus a result of a multitude of forces originating from not only from the molecules directly involved, but due to the crowded nature of the cell's environment, also other molecules in the surroundings. Fortunately, keeping track of the exact trajectories of all these particles is not actually needed in order to extract useful (average) measures of a chemical reaction's outcome. We have entered the realm of statistical mechanics, in which we want to know what is most likely to happen when repeating a chemical reaction many times (as is typical). If molecule A reacts with B to form species C, what is the concentration of molecule C after a time t? Or say species A is part of multiple chemical pathways and is capable of reacting either with species B_1 or with B_2, which is more likely to happen sooner? We shall cover the most important techniques used to tackle such questions.

As an example, let us take a simplified view of an RGN interacting with its target substrate. The top panel of figure **??**A shows a chemical reaction in which a target bound RGN cleaves its substrate. Below is drawn what is called the free-energy landscape for this reaction. Any possible set of positions of the RGN, target (or parts thereof) – together this will be referred to as our 'system' – is summarized as one configuration along the horizontal axis in the diagram – essentially starting from an unbound configuration on the left to a cleaved product on the right. The only number we keep track of is what is called the system's free-energy

($F = E–TS$) – a combination of its internal energy (E) and conformational entropy (S) at a fixed temperature (T).

As far as our chemical reaction goes, we are not interested in any intermediate positional configuration in which the substrate is not cleaved yet or the substrate is not bound yet. We shall discuss below that a lower free-energy describes a more likely configuration or 'state'. Hence, we represent the reactants (RGN is bound to substrate) and products (substrate gets cleaved) as local (or global) minima in the free-energy landscape. Completing the reaction requires the system to first overcome an energetic barrier – the amount of ΔF - to take it over the local maximum called the transition state (T). In this thesis we used what is called 'kinetic modeling', in which we assume the time for any single reaction (one arrow in the diagram) to get completed to be exponentially distributed (**Figure ??**B and C). Using $p(t)$ to denote the probability of not having completed the reaction of figure **??** before time t:

$$\phi(t) = ke^{-kt} \tag{1.1}$$

$$1 - p(t) = \int_0^t \phi(t)\mathrm{d}t = 1 - e^{-kt} \tag{1.2}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -kp(t) \tag{1.3}$$

The inverse average time of the reaction, or reaction rate, k, is related to the (free-)energy barrier of the reaction through the Arrhenius equation:

$$k \propto e^{-\Delta F} \tag{1.4}$$

Throughout this thesis, we shall measure all energies in units of the thermal energy k_BT. A slightly more complicated reaction is one in which the RGN toggles between being unbound

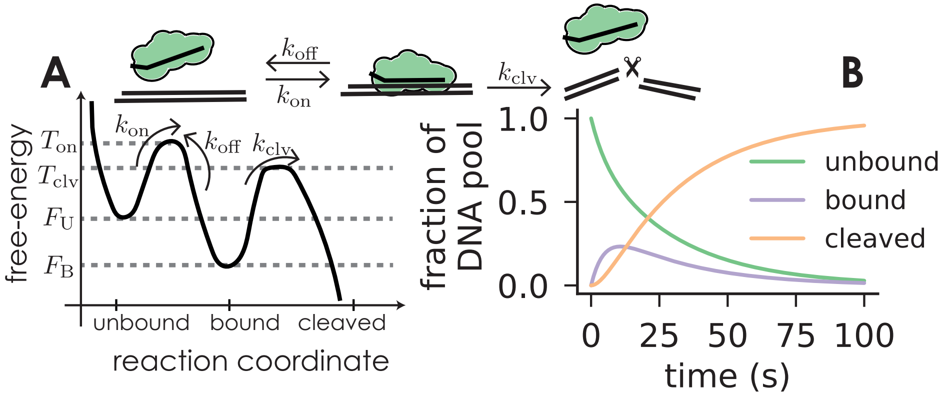


Figure 1.11: The RGN binds its substrate at a rate k_{on}. Before cleaving with rate k_{clv}, the RGN can unbind at a rate k_{off}. **(A)** Free-energy landscape. Stable states (minima) are denoted by F's, while transition states between two configurations are indicated by T's. **(B)** Solution to Master equation tracks populations of all the three states over time.

and bound to its substrate before it can cleave it (**Figure ??**). Its corresponding free-energy landscape is shown in figure **??**A. By extension of the previous example, every step within this reaction scheme is characterized by a minima and a set of transition barriers separating it from subsequent steps. The Arrhenius equation relates these barriers to reaction rates. How do we now track the fraction of cleaved DNA? First note that all RGN and substrate molecules must belong to one of the species described in the chemical reaction pathway. Their relative fractions, or the probability that any of the molecules belongs to a given species, can vary over time, but the total is conserved:

$$p_{\mathrm{ub}}(t) + p_{\mathrm{bnd}}(t) + p_{\mathrm{clv}}(t) = 1 \;\; \forall t \tag{1.5}$$

In this example the number of unbound molecules at a time t decreases by unbound molecules binding to a substrate. On average, every k_{on}^{-1} seconds an unbound molecule binds. For this to happen at time t, there must be an unbound molecule available at time t to start with. Hence, the rate of change of the unbound population decreases by a factor of $p_{\mathrm{ub}}(t) \times k_{\mathrm{on}}$. Similarly, when a bound molecule rejects its substrate, the fraction of unbound molecules increases. Taken together, the set of differential equations describing the time evolution of all of the different populations, termed the set of Master Equations, are

$$\frac{\mathrm{d}p_{\mathrm{ub}}}{\mathrm{d}t} = -k_{\mathrm{on}} p_{\mathrm{ub}}(t) + k_{\mathrm{off}} p_{\mathrm{bnd}}(t) \tag{1.6}$$

$$\frac{\mathrm{d}p_{\mathrm{bnd}}}{\mathrm{d}t} = +k_{\mathrm{on}} p_{\mathrm{ub}}(t) - (k_{\mathrm{off}} + k_{\mathrm{clv}}) p_{\mathrm{bnd}}(t) \tag{1.7}$$

$$\frac{\mathrm{d}p_{\mathrm{clv}}}{\mathrm{d}t} = +k_{\mathrm{clv}} p_{\mathrm{bnd}}(t) \tag{1.8}$$

Commonly, one re-writes it in matrix-vector form ($\vec{P}(t) = [p_{\mathrm{ub}}(t), p_{\mathrm{bnd}}(t), p_{\mathrm{clv}}(t)]^T$):

$$\frac{\mathrm{d}\vec{P}}{\mathrm{d}t} = M\vec{P}(t)\,,\; M = \begin{pmatrix} -k_{\mathrm{on}} & k_{\mathrm{off}} & 0 \\ +k_{\mathrm{on}} & -(k_{\mathrm{off}} + k_{\mathrm{clv}}) & 0 \\ 0 & k_{\mathrm{clv}} & 0 \end{pmatrix} \tag{1.9}$$

The solution for this particular problem is plotted in figure **??**B. In general, solving the Master Equations gives us access to all concentrations of reactants and products for any particular reaction pathway.

1.7.2. When reactions are fast: Equilibrium Thermodynamics

The reactions described above eventually become irreversible - after the RGN cuts its substrate there is no way back (see **Figure ??-??**). However, the sub-process of substrate binding is reversible. If the binding and unbinding happen much faster than cleaving ($k_{\mathrm{off}}, k_{\mathrm{on}} \gg k_{\mathrm{clv}}$, see **Figure ??**A), a (local) equilibrium between bound and unbound states may be reached prior to cleaving. In other words, the bound and unbound states will essentially evolve together as if it being a closed system. After the two have saturated (equilibrated) the fraction of cleaved DNA is still set by the rate k_{clv} and the now locally equilibrated fraction of bound molecules (equation **??**).

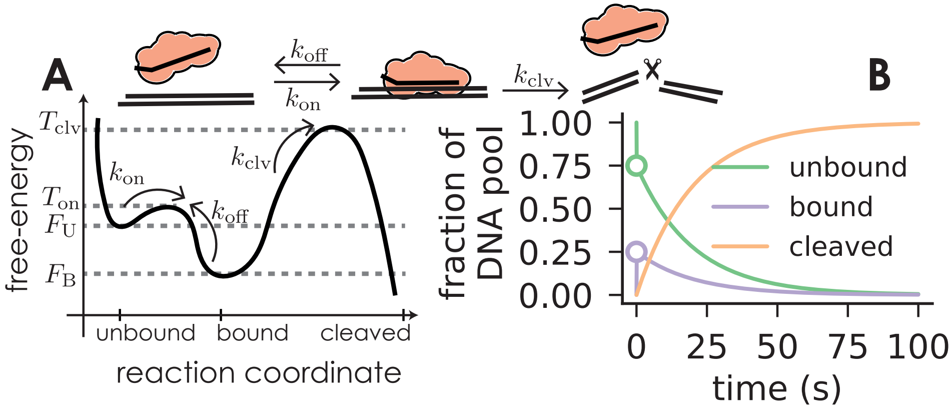


Figure 1.12: Same reaction as in **Figure ??**, but now with much higher rates of binding and unbinding. **(A)** Free-energy landscape. **(B)** Solution to Master equation tracks populations of all the three states over time. Dots indicate equillibrium fractions of bound/unbound DNA calculated using the Detailed balance condition of eq (**??**).

This trick of separating timescales can greatly simplify the solution to the Master Equations, since the populations in the equilibrated states are known to satisfy the Boltzmann distribution, which is time independent.

$$p^{\mathrm{EQ}}_{\mathrm{ub,b}} = \frac{e^{-F_{\mathrm{ub,b}}}}{e^{-F_{\mathrm{ub}}} + e^{-F_{\mathrm{b}}}} \equiv \frac{e^{-F_{\mathrm{ub,b}}}}{Z} \tag{1.10}$$

with Z commonly referred to as the system's partition function. Allowing for processes to locally equilibrate serves as a means to account for the relevant 'slow' reaction involving k_{clv} and ignore any temporal contributions of the very short times (k_{on}^{-1} and k_{off}^{-1}). Yet, we did not lose the information that an unbound molecule must first bind before it is able to cleave - as the stability of the bound state decreases, so does the fraction of bound molecules.

In case of the molecule being completely incapable of cleaving ($k_{\mathrm{clv}} \rightarrow 0$ or equivalently $T_{\mathrm{clv}} \rightarrow \infty$) the bound and unbound states form a completely closed system. Hence, a (global) equilibrium will be reached eventually. At the same time, the master equation approach to determine concentrations of either bound or unbound molecules at shorter times is still valid. How do we choose the set of rates in the Master Equation to ensure that the resulting probabilities approach the according values determined by the Boltzmann distribution? When equilibrated, the probability is stationary, which written in terms of the Master equations reads as follows:

$$p^{\mathrm{EQ}}_{\mathrm{ub}} k_{\mathrm{on}} = p^{\mathrm{EQ}}_{\mathrm{bnd}} k_{\mathrm{off}} \tag{1.11}$$

Equation **??** says that the flow of probability out of the (un)bound state equals the flow into it. Hence, setting the rates according to this 'Detailed balance condition'

$$k_{\mathrm{off}} = k_{\mathrm{on}} \frac{p^{\mathrm{EQ}}_{\mathrm{ub}}}{p^{\mathrm{EQ}}_{\mathrm{bnd}}} = k_{\mathrm{on}} e^{-(F_{\mathrm{ub}} - F_{\mathrm{b}})} \tag{1.12}$$

guarantees that the probabilities will approach their appropriate Boltzmann weights:

$$\lim_{t\to\infty} p_{\mathrm{i}}(t) = p_{\mathrm{i}}^{\mathrm{EQ}} = \frac{e^{F_{\mathrm{i}}}}{Z} \; \forall i \in [\mathrm{bnd}, \mathrm{ub}] \tag{1.13}$$

In figure **??**B, the two dots shown are the equilibrium fractions calculated using (the inverse of) equation **??**. For the more involved reaction pathways considered later in this thesis, the detailed balance condition is applied for every pair of adjacent states i and j: $k_{i\to j}/k_{j\to i} = e^{\Delta F_{\mathrm{ij}}}$.

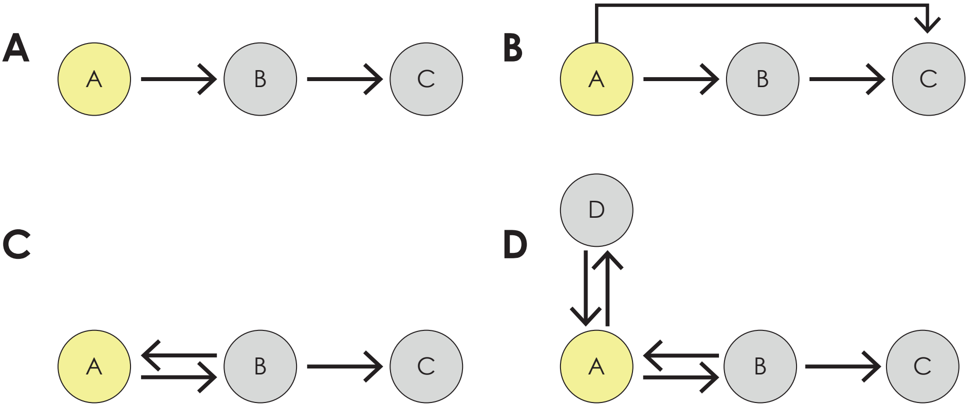


Figure 1.13: Four examples of first passage problems. In each of the figures **(A)**- **(D)** we seek the first time we arrive at node C, starting from node A. The probability (density) that this occurs at time t is denoted by $\Psi_{AC}(t)$.

1.7.3. First Passage Problems of Continuous Time Random Walks

Without solving the Master equations, we can still determine the average time needed to complete a chemical reaction or its most likely outcome. To do such we pretend the chemical reaction is actually a random walk on a lattice with each intermediate representing a node. The walker takes a step on the lattice by completing a single reaction, thereby taking a step along an arrow shown in the diagram. A convenient way of approaching these problems will be to view them as little 'board games'. Walking on the board is done by hopping from one node to another, one at the time and only along a direction indicated by an arrow. Here we focus on some 'rules of the game'.

The first important rule is that we only record the time in between transitions. Transitions themselves happen instantaneously. It is as if we are playing a game of 'speed chess' in which we record the times it takes to decide what moves to make, not the time needed to actually move the piece across the board. More formally, when one considers the movement of a body on an interval $x \in [a, b]$, then a simple question one may ask is:*"What is the time at which the particle passes the boundary at a or b for the first time?"*. This time is called the 'first passage time'. One may also ask: *"What is the probability that the first passage time at boundary a equals $t = t_a$?"*. This will be referred to as the first passage probability. Let $\Psi(t)\mathrm{d}t$ denote the probability that the first passage time lies within

$[t, t+\mathrm{d}t]$. Hence, $\Psi(t)$ is the 'first passage probability density'. Within the context of chemical reactions, the first passage time indicates when a reaction gets completed for the first time. Hence, as mentioned, our random walk will take place on a discretized spatial lattice. Time, however, remains a continuous variable. In each of the following examples, we will be after the first passage at node C, starting from node A ($\Psi(t)_{AC}$).

As an example, consider the 'board game' shown in figure **??**A. Starting from node A, our next move can only take us to a node that neighbors A and for which there is an arrow pointing in the designated direction. In this case, we have little choice but walking to node B. Let $\phi_{XY}(t)$'s denote the probability densities of reaction times for individual reactions - representing exponential distributions (see equation **??**) - making a step from X to Y. What is the probability that one arrives at node C at time t? Given node A is not directly connected to node C, all possible paths that bring us to node C must have first brought us to node B at some earlier time $\tau < t$.

$$\Psi_{AC}(t) = \int_0^\infty \phi_{AB}(\tau)\phi_{BC}(t-\tau)\mathrm{d}\tau \tag{1.14}$$

The above integral reflects that we must sum over all possible ways of ending up at C via node A - increasing for an increasing number of ways of getting to the designation. In this case it entails summing over all times at which we arrived at the intermediate node B, resulting in the convolution of $\phi_{AB}(t)$ and $\phi_{BC}(t)$. If we instead use Laplace transforms of the probability densities -

$$\Psi_{AC}(s) = \mathcal{L}\{\Psi_{AC}(t)\} = \int_0^\infty \Psi_{AC}(t)e^{-st}\mathrm{d}t \tag{1.15}$$

- such a convolution turns into a simple product in s-space:

$$\begin{aligned}
\Psi_{AC}(s) &= \mathcal{L}\left\{\int_0^\infty \phi_{AB}(\tau)\phi_{BC}(t-\tau)\mathrm{d}\tau\right\} \\
&= \int_0^\infty \int_0^\infty \phi_{AB}(\tau)\phi_{BC}(t-\tau)\mathrm{d}\tau e^{-st}\mathrm{d}t \\
&= \int_0^\infty \int_0^\infty \phi_{AB}(\tau)\phi_{BC}(t-\tau)e^{-st}\mathrm{d}\tau\mathrm{d}t \\
&\equiv \int_0^\infty \int_0^\infty \phi_{AB}(\tau)\phi_{BC}(u)e^{-s(u+\tau)}\mathrm{d}\tau\mathrm{d}u \\
&= \int_0^\infty \phi_{AB}(\tau)e^{-s\tau}\mathrm{d}\tau \int_0^\infty \phi_{BC}(u)e^{-su}\mathrm{d}u \\
&= \phi_{AB}(s) \times \phi_{BC}(s)
\end{aligned} \tag{1.16}$$

The Laplace transform is also a linear operator, which we shall put to practice in the example shown in figure **??**B. In this example there are two distinct types of paths that lead from A to C. We can walk directly from A to C (ϕ_{AC}) or use node B as an intermediate. Summing over the distinct paths equals summing over the corresponding Laplace transforms.

$$\Psi_{AC}(s) = \phi_{AB}(s)\phi_{BC}(s) + \phi_{AC}(s) \tag{1.17}$$

Only for a select set of problems it is possible to directly invert the Laplace transform. Fortunately, this is not needed in order to obtain the mean first passage time at C starting from A. For this, consider the derivate of $\Psi(s)$, evaluated at $s = 0$.

$$\begin{aligned}\left(\frac{\mathrm{d}\Psi_{AC}}{\mathrm{d}s}\right)_{s=0} &= \left(\int_0^\infty \Psi_{AC}(t)\frac{\mathrm{d}e^{-st}}{\mathrm{d}s}\mathrm{d}t\right)_{s=0} \\ &= \int_0^\infty -t\Psi_{AC}(t)e^0\mathrm{d}t \\ &\equiv -\langle t\rangle\end{aligned} \tag{1.18}$$

In general, the n^{th} order moment of the first passage time - the first moment is called the mean - is obtained by taking the n^{th} order derivative of the Laplace transform. The function $\Psi(s)$ is therefore also referred to as the moment generating function.

$$\langle t^n\rangle = (-1)^n\left(\frac{\mathrm{d}^n\Psi_{AC}}{\mathrm{d}s^n}\right)_{s=0} \tag{1.19}$$

The 0^{th}order moment has a special interpretation,

$$P \equiv \Psi_{AC}(0) = \int_0^\infty \Psi_{AC}(t)\mathrm{d}t \tag{1.20}$$

It equals the probability of completing the specified reaction first. Later in this thesis we will use exactly this probability to determine if a bound RGN will cleave before it unbinds. Note that for all the board games shown in figure **??**, this probability must equal one as node C is the only final product possible.

There are two more ’rules of the game’ that have come in extremely handy in later chapters. First consider the example of figure **??**C. The board reveals that node C cannot be reached within a single step. We must walk to node B first. However, unlike in figure A there are many ways in which we can get to node C (for the first time). After walking to node B, we can decide to walk back to point A, then back to B and finally walk to C. As a matter of fact, we can decide to walk back and forth between A and B as often as we want as long as we end by taking a step from A to B and one from B to C. Using both the convolution property and the linearity of the Laplace transform we find

$$\begin{aligned}\Psi_{AC} &= \left[1 + (\phi_{AB}\phi_{BA}) + (\phi_{AB}\phi_{BA})^2 + (\phi_{AB}\phi_{BA})^3 + ...\right]\phi_{AB}\phi_{BC} \\ &= \sum_{n=0}^{\infty}(\phi_{AB}\phi_{BA})^n]\phi_{AB}\phi_{BC} \\ &= \frac{\phi_{AB}\phi_{BC}}{1-\phi_{AB}\phi_{BA}}\end{aligned} \tag{1.21}$$

The last line follows from recognizing the geometric series.

Let us turn to one final example, figure **??**D. Before to walking to C, we may walk back and forth between A and B several times. Similarly, we are allowed to walk back and forth between A and D as often as we like. We can even walk along the path A-D-A-B-A-D-A-B-C, or any other combination in which we toggle between the nodes A,B and D before making it to C. The previous example demonstrated that dealing with a single 'two-way-arrow' - one reversible reaction - results in a sum of terms of the form $\phi_{AB}\phi_{BA}$ or $\phi_{AD}\phi_{DA}$. At first glance, one may expect the solution to this problem to be $\Psi = \sum_n (\phi_{AB}\phi_{BA})^n \times \sum_m (\phi_{AD}\phi_{DA})^m$. Although any valid path from A to C is indeed represented by a term in the sum, we are not accounting for the fact that many paths are now represented by one and the same contribution. A first passage for which there are more paths leading to it should become more likely. We are therefore still missing a combinatorial factor describing the number of ways the pairs for $\phi_{AB}\phi_{BA}$ and $\phi_{AD}\phi_{DA}$ can commute. Instead of doing explicit combinatorics, counting every possible path by hand, we will still approach the problem in a similar fashion as we did in the previous example. Before, we characterized a particular path by the number of times one stepped back and forth, using node B in figure **??**C. Let us do the same, now using the board game of figure D. Say we walked back and forth twice, without knowing whether we used node B or D any of the following paths could have been taken:

- Use node B twice, walk A-B-A-B-A(-B-C): $(\phi_{AB}\phi_{BA})^2$
- Use node D twice, walk A-D-A-D-A(-B-C):$(\phi_{AD}\phi_{DA})^2$
- First use B, then use D. walk A-B-A-D-A(-B-C): $\phi_{AB}\phi_{BA} \times \phi_{AD}\phi_{DA}$
- First use D, then use B. walk A-D-A-B-A(-B-C): $\phi_{AD}\phi_{DA} \times \phi_{AB}\phi_{BA}$

Taken together, Ψ_{AC}, must gather a term equal to:

$$(\phi_{AB}\phi_{BA})^2 + 2(\phi_{AD}\phi_{DA}\phi_{AB}\phi_{BA}) + (\phi_{AD}\phi_{DA})^2 = (\phi_{AD}\phi_{DA} + \phi_{AB}\phi_{BA})^2 \quad (1.22)$$

Generalizing this example shows that walking back and forth a total of n times contributes a term of $(\phi_{AD}\phi_{DA} + \phi_{AB}\phi_{BA})^n$ to $\Psi_{AC}(s)$.

$$\Psi_{AC}(s) = \sum_n (\phi_{AD}\phi_{DA} + \phi_{AB}\phi_{BA})^n \times \phi_{AB}\phi_{BC} \quad (1.23)$$

1.7.4. Decision making: The 'splitting probability'

As a final piece of theory - tying together the first passage problems and the master equation approaches - consider a bound RGN that can partake in one of two irreversible reactions: unbinding (ignore (re-)binding) at a rate k_{off} and cleavage at a rate k_{clv}. When we speak of the 'total outgoing rate' from the bound state we are referring to $k = k_{\text{ub}} + k_{\text{clv}}$. Note that the conditional waiting time(s) are distributed as follows:

$$\phi_i = k_{\text{i}} e^{\sum_x k_x} \; \forall i, x \in [\text{ub}, \text{clv}] \quad (1.24)$$

This is the generalisation of equation **??**. Hence, if one tracks the number of bound molecules, this number will decrease exponentially at a total rate of k, irrespective if a molecule cleaves or rejects the substrate. To know if the RGN is more likely to cleave before it unbinds, or vice versa, we take a look at the zeroth order moment of its Laplace transform.

$$\phi_i(s) = \frac{k_i}{s + \sum_x k_x} \tag{1.25}$$

For sake of illustration, we use the Laplace transform even though the corresponding integral in the temporal domain is easy to compute. Taking any of these two approaches,

$$P_i = \frac{k_\mathrm{i}}{\sum_x k_x} \tag{1.26}$$

This is commonly referred to as the 'splitting probability' of reaction path i.

1.7.5. Connection to experimental data

Throughout this thesis validating our model predictions against experimental data forms a crucial part of the research presented. Bulk biochemical assays may report on the fraction of cleaved molecules after some fixed time. We can either use the master equation to obtain this same quatity, or work within limmits wherin it should be well approximated by the ratio in reaction rates for the different off-target molecules (inverse average times), or the (splitting) probability for cleaving. Other assays use fluorescent labels to track populations of substrates and RGNs over time, thereby directly reporting on the solution to the corresponding Master Equation. Finally, single-molecule experiments enable one to track individual guide-loaded RGN complexes, which allows one to directly measure (mean) first passage times or the time distributions $\phi(t)$ - or the total distribution of $\Psi(t)$ in case of a more complex chemical pathway.

1.8. In this thesis

This thesis is an account of modeling efforts aimed towards understanding the kinetics underlying (off-)targeting by RNA/DNA guided nucleic acid effector complexes.

Part I: Target recognition and off-target prediction quantifies what types of off-targets lead to cleavage before rejection, with a particular focus on the position of mismatches within the guide-target hybrid.
Chapter ?? introduces a kinetic model for the off-target binding and cleavage by CRISPR-Cas, Argonaute, and similar RNA guided nucleases (RGNs). Previous literature revealed such RGNs bind their substrate and aid the formation of the guide-target hybrid in sequential fashion. Using a minimalistic view of target recognition, we say the addition of a match to the hybrid is energetically (and kinetically) favorable, whereas a mismatch biases the system towards rejection of the off-target. Working out the mathematics purely dictated by the targeting process being sequential, allows us to give a physical explanation for a multitude of empirically derived 'off-targeting rules' – a set of 'rules of thumb' experimenters adhere to when designing their RGN-based assay.
In **Chapter ??** we built upon this model by expanding the parameterization to include position dependent (mis-)match biases. Using a series of high-throughput biophysical datasets

we elucidate the free-energy landscape that underlies *Streptococcus pyogenes* Cas9 (spCas9) target recognition. Previous reports showed catalytically 'dead' Cas9 (dCas9) binds many more (genomic) off-targets than active Cas9 cleaves. The presented free-energy landscape not only unifies those observations, but explains exactly what off-targets lead to stable binding, apparently without getting cut. In particular, our model allows one to calculate how much off-target binding by dCas9 or cleavage by Cas9 is to be expected given the nuclease concentration and reaction time used in an experiment. Finally, the free-energy landscape further reveals Cas9's major conformational change, in which it repositions its nuclease domains to enable cleavage, directly couples to the entire hybrid formation process.

Thus far, we have been treating the selection/rejection of isolated off-targets. **Part II: Target search** focuses on how sequence specific binding proteins locate their cognate target site amongst a pool of potential off-targets. Apart from diffusing through solution until the protein randomly collides with a target, proteins are found to enhance their reaction rates by binding non-specifically and diffusing laterally along the DNA/RNA.
Chapter ?? uses the example of hAgo2 to review existing target search literature and hypothesizes that a coupling of the protein's structural changes to the hybrid formation – much like the kind found for spCas9 in **Chapter ??** – balances search time and specificity.

Typically, the target search is further complicated as large portions of cellular RNA/DNA are occupied by other proteins. Moreover, the RNA/DNA is highly compacted, adopting a conformation that severely deviates from being linear, even on the scale of the searching protein. In **Chapter ??** we used a prokaryotic Argonaute as a model system to investigate if and how lateral diffusion can proceed in the presence of either structural or protein obstacles. The presented single-molecule FRET experiments (a collaboration with T.J.Cui from the lab of dr. Chirlmin Joo) demonstrate cbAgo can bypass both a secondary DNA structure (a 'Y-fork') and a bound protein - covering DNA sites at (nearly) the same rates as on bare DNA. Using kinetic modeling allowed us to further demonstrate that the secondary structure does not hinder the lateral sliding motion, while the bulkier protein barrier does - necessitating some form of dissociation from the DNA in order to 'skip' over the obstacle in order to proceed searching.
Motivated by these observations, we ask whether a laterally diffusing protein must interrogate all (off-)targets along its path in **Chapter ??**. We set up a rather generic model that allows for the protein to interrogate only a fraction of all sites enclosed within its lateral excursion. Using single-molecule FRET experiments performed on both a bacterial Ago and hAgo2, our model shows both systems indeed only interrogate a relatively small fraction of all DNA/RNA sites. Surprisingly, despite essentially "being blind" to a significant portion of the target pool, we show how this can actually help to find the cognate site faster.

References

[] E. S. L. Lander, L. M. Birren, B. Nusbaum, C. Zody, M. C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. Funke, R. Gage, D. Harris, K. Heaford, A. Howland, J. Kann, L. Lehoczky, J. LeVine, R. McEwan, and P. McKernanKevin., *Initial sequencing and analysis of the human genome,* Nature **409**, 860 (2001).

[] Ø. Bergh, K. Y. Børsheim, G. Bratbak, and M. Heldal, *High abundance of viruses found in aquatic environments,* Nature **340**, 467 (1989).

[] M. Breitbart and F. Rohwer, *Here a virus, there a virus, everywhere the same virus?* Trends in Microbiology **13**, 278 (2005).

[] A. Stern and R. Sorek, *The phage-host arms race: Shaping the evolution of microbes,* BioEssays **33**, 43 (2011).

[] R. Sorek, C. M. Lawrence, and B. Wiedenheft, *CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea,* Annual Review of Biochemistry **82**, 237 (2013).

[] B. Wiedenheft, S. H. Sternberg, and J. A. Doudna, *RNA-guided genetic silencing systems in bacteria and archaea,* Nature **482**, 331 (2012), arXiv:37 .

[] F. J. Mojica, C. Díez-Villaseñor, J. García-Martínez, and E. Soria, *Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements,* Journal of Molecular Evolution **60**, 174 (2005).

[] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath, *CRISPR provides acquired resistance against viruses in prokaryotes,* Science **315**, 1709 (2007), arXiv:arXiv:1011.1669v3 .

[] J. K. Nuñez, P. J. Kranzusch, J. Noeske, A. V. Wright, C. W. Davies, and J. A. Doudna, *Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity,* Nature Structural and Molecular Biology **21**, 528 (2014).

[] S. J. Brouns, M. M. Jore, M. Lundgren, E. R. Westra, R. J. Slijkhuis, A. P. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin, and J. Van Der Oost, *Small Crispr Rnas Guide Antiviral Defense in Prokaryotes,* Science **321**, 960 (2008), arXiv:20 .

[] G. Gasiunas, R. Barrangou, P. Horvath, and V. Siksnys, *Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria,* Proceedings of the National Academy of Sciences **109**, E2579 (2012), arXiv:arXiv:1408.1149 .

[] J. E. Garneau, M. È. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadán, and S. Moineau, *The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA,* Nature **468**, 67 (2010).

[] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, *A Programmable Dual-RNA – Guided,* Science **337**, 816 (2012), arXiv:38 .

[] R. T. Leenay, K. R. Maksimchuk, R. A. Slotkowski, R. N. Agrawal, A. A. Gomaa, A. E. Briner, R. Barrangou, and C. L. Beisel, *Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems,* Molecular Cell **62**, 137 (2016).

[] D. G. Sashital, B. Wiedenheft, and J. A. Doudna, *Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System,* Molecular Cell **46**, 606 (2012).

[] E. R. Westra, E. Semenova, K. A. Datsenko, R. N. Jackson, B. Wiedenheft, K. Severinov, and S. J. Brouns, *Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition,* PLoS Genetics **9** (2013), 10.1371/journal.pgen.1003742.

[] K. S. Makarova, Y. I. Wolf, O. S. Alkhnbashi, F. Costa, S. A. Shah, S. J. Saunders, R. Barrangou, S. J. Brouns, E. Charpentier, D. H. Haft, P. Horvath, S. Moineau, F. J. Mojica, R. M. Terns, M. P. Terns, M. F. White, A. F. Yakunin, R. A. Garrett, J. Van Der Oost, R. Backofen, and E. V. Koonin, *An updated evolutionary classification of CRISPR-Cas systems,* Nature Reviews Microbiology **13**, 722 (2015), arXiv:9809069v1 [arXiv:gr-qc] .

[] T. Sinkunas, G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath, and V. Siksnys, *Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system,* EMBO Journal **30**, 1335 (2011).

[] S. N. Kieper, C. Almendros, J. Behler, R. E. McKenzie, F. L. Nobrega, A. C. Haagsma, J. N. Vink, W. R. Hess, and S. J. Brouns, *Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation,* Cell Reports **22**, 3377 (2018).

[] L. B. Harrington, J. S. Chen, E. Ma, I. P. Witte, J. C. Cofsky, J. A. Doudna, D. Burstein, J. F. Banfield, D. Paez-Espino, and N. C. Kyrpides, *Programmed DNA destruction by miniature CRISPR-Cas14 enzymes,* Science **362**, 839 (2018).

[] R. C. Friedman, K. K. H. Farh, C. B. Burge, and D. P. Bartel, *Most mammalian mRNAs are conserved targets of microRNAs,* Genome Research **19**, 92 (2009).

[] D. P. Bartel, *MicroRNAs: Target Recognition and Regulatory Functions,* Cell **136**, 215 (2009), arXiv:0208024 [gr-qc] .

[] G. Hutvagner and M. J. Simard, *Argonaute pro-*

teins: Key players in RNA silencing, Nature Reviews Molecular Cell Biology **9**, 22 (2008).

[] T. Kawamata and Y. Tomari, *Making RISC,* Trends in Biochemical Sciences **35**, 368 (2010).

[] G. Meister, *Argonaute proteins: Functional insights and emerging roles,* Nature Reviews Genetics **14**, 447 (2013).

[] A. Z. Fire, X. SiQun, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans,* Nature **391**, 806 (1998).

[] *https://www.nobelprize.org/prizes/medicine /2006/press-release,* .

[] D. C. Swarts, K. Makarova, Y. Wang, K. Nakanishi, R. F. Ketting, E. V. Koonin, D. J. Patel, and J. Van Der Oost, *The evolutionary journey of Argonaute proteins,* Nature Structural and Molecular Biology **21**, 743 (2014).

[] J. W. Hegge, D. C. Swarts, S. D. Chandradoss, T. J. Cui, J. Kneppers, M. Jinek, C. Joo, and J. van der Oost, *DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute,* Nucleic Acids Research **47**, 5809 (2019).

[] E. Kaya, K. W. Doxzen, K. R. Knoll, R. C. Wilson, S. C. Strutt, P. J. Kranzusch, and J. A. Doudna, *A bacterial Argonaute with noncanonical guide RNA specificity,* Proceedings of the National Academy of Sciences **113**, 4057 (2016), arXiv:arXiv:1408.1149 .

[] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, *Multiplex genome engineering using CRISPR/Cas systems,* Science **339**, 819 (2013), arXiv:20 .

[] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church, *RNA-guided human genome engineering via Cas9,* Science **339**, 823 (2013).

[] E. Deltcheva, K. Chylinski, C. M. Sharma, K. Gonzales, Y. Chao, Z. A. Pirzada, M. R. Eckert, J. Vogel, and E. Charpentier, *CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III,* Nature **471**, 602 (2011).

[] J. Shi, H. Gao, H. Wang, H. R. Lafitte, R. L. Archibald, M. Yang, S. M. Hakimi, H. Mo, and J. E. Habben, *ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions,* Plant Biotechnology Journal **15**, 207 (2017).

[] D. Bikard, C. W. Euler, W. Jiang, P. M. Nussenzweig, G. W. Goldberg, X. Duportet, V. A. Fischetti, and L. A. Marraffini, *Exploiting CRISPR-cas nucleases to produce sequence-specific antimicrobials,* Nature Biotechnology **32**, 1146 (2014), arXiv:NIHMS150003 .

[] L. Amoasii, H. Li, E. Sanchez-Ortiz, A. Mireault, D. Caballero, R. Bassel-Duby, E. N. Olson, J. C. Hildyard, R. Harron, C. Massey, R. J. Piercy, T.-R. Stathopoulou, and J. M. Shelton, *Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy,* Science **362**, 86 (2018).

[] B. Zetsche, J. S. Gootenberg, O. O. Abudayyeh, I. M. Slaymaker, K. S. Makarova, P. Essletzbichler, S. E. Volz, J. Joung, J. Van Der Oost, A. Regev, E. V. Koonin, and F. Zhang, *Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System,* Cell **163**, 759 (2015), NIHMS150003 .

[] L. A. Gilbert, M. H. Larson, L. Morsut, Z. Liu, G. A. Brar, S. E. Torres, N. Stern-Ginossar, O. Brandman, E. H. Whitehead, J. A. Doudna, W. A. Lim, J. S. Weissman, and L. S. Qi, *CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes,* Cell **154**, 442 (2013).

[] B. Chen, L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G. W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, and B. Huang, *Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system,* Cell **155**, 1479 (2013).

[] J. S. Gootenberg, O. O. Abudayyeh, J. W. Lee, P. Essletzbichler, A. J. Dy, J. Joung, V. Verdine, N. Donghia, N. M. Daringer, C. A. Freije, C. Myhrvold, R. P. Bhattacharyya, J. Livny, A. Regev, E. V. Koonin, D. T. Hung, P. C. Sabeti, J. J. Collins, and F. Zhang, *Nucleic acid detection with CRISPR-Cas13a/C2c2.* Science (New York, N.Y.) **356**, 438 (2017), arXiv:15334406 .

[] W. K. Spoelstra, J. M. Jacques, F. L. Nobrega, A. C. Haagsma, M. Dogterom, T. Idema, S. J. Brouns, and L. Reese, *CRISPR-based DNA and RNA detection with liquid phase separation,* Bioarxiv , 1 (2018).

[] X. Wang, E. Xiong, T. Tian, M. Cheng, W. Lin, and J. Sun, *CASLFA : CRISPR / Cas9-mediated lateral flow nucleic acid assay,* Bioarxiv (2019).

[] D. Kim, K. Luk, S. A. Wolfe, and J.-S. Kim, *Evaluating and Enhancing Target Specificity of Gene-Editing Nucleases and Deaminases,* Annual Review of Biochemistry , 1 (2019).

[] S. Q. Tsai and J. K. Joung, *Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases,* Nature Reviews Genetics **17**, 300 (2016).

[] N. Amrani, X. D. Gao, P. Liu, A. Edraki, A. Mir, R. Ibraheim, A. Gupta, K. E. Sasaki, T. Wu, P. D. Donohoue, A. H. Settle, A. M. Lied, K. McGovern, C. K. Fuller, P. Cameron, T. G. Fazzio, L. J. Zhu, S. A. Wolfe, and E. J. Sontheimer, *NmeCas9 is an intrinsically high-fidelity genome-editing platform Jin-Soo Kim,* Genome Biology **19**, 1 (2018).

[] B. P. Kleinstiver, S. Q. Tsai, M. S. Prew, N. T. Nguyen, M. M. Welch, J. M. Lopez, Z. R. McCaw, M. J. Aryee, and J. K. Joung, *Genome-wide* *specificities of CRISPR-Cas Cpf1 nucleases in human cells,* Nature Biotechnology **34**, 869 (2016), arXiv:15334406 .

[] H. Wang, M. La Russa, and L. S. Qi, *CRISPR/Cas9 in Genome Editing and Beyond,* Annual Review of Biochemistry **85**, 227 (2016).

[] Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio, and J. K. Joung, *Improving CRISPR-Cas nuclease specificity using truncated guide RNAs,* Nature Biotechnology **32**, 279 (2014), arXiv:29 .

[] J. S. Chen, Y. S. Dagdas, B. P. Kleinstiver, M. M. Welch, A. A. Sousa, L. B. Harrington, S. H. Sternberg, J. K. Joung, A. Yildiz, and J. A. Doudna, *Enhanced proofreading governs CRISPR-Cas9 targeting accuracy,* Nature **550**, 407 (2017).

[] B. P. Kleinstiver, V. Pattanayak, M. S. Prew, S. Q. Tsai, N. T. Nguyen, Z. Zheng, and J. K. Joung, *High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects,* Nature **529**, 490 (2016), arXiv:9605103 [cs] .

[] I. M. Slaymaker, L. Gao, B. Zetsche, D. A. Scott, W. X. Yan, and F. Zhang, *Rationally engineered Cas9 nucleases with improved specificity,* Science **351**, 84 (2016), arXiv:NIHMS150003 .

[] S. Bae, J. Park, and J. S. Kim, *Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases,* Bioinformatics **30**, 1473 (2014).

[] K. Labun, T. G. Montague, J. A. Gagnon, S. B. Thyme, and E. Valen, *CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering,* Nucleic acids research **44**, W272 (2016).

[] F. Heigwer, G. Kerr, and M. Boutros, *E-CRISP: Fast CRISPR target site identification,* Nature Methods **11**, 122 (2014).

[] P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, T. J. Cradick, L. A. Marraffini, G. Bao, and F. Zhang, *DNA targeting specificity of RNA-guided Cas9 nucleases,* Nature Biotechnology **31**, 827 (2013), arXiv:NIHMS150003 .

[] M. Stemmer, T. Thumberger, M. Del Sol Keyer, J. Wittbrodt, and J. L. Mateo, *CCTop: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool,* PLoS ONE **10**, 1 (2015).

[] J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten, and D. E. Root, *Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9,* Nature Biotechnology **34**, 184 (2016), arXiv:15334406 .

[] G. Chuai, H. Ma, J. Yan, M. Chen, N. Hong, D. Xue, C. Zhou, C. Zhu, K. Chen, B. Duan, F. Gu, S. Qu, D. Huang, J. Wei, and Q. Liu, *DeepCRISPR: Optimized CRISPR guide RNA design by deep learning,* Genome Biology **19**, 1 (2018).

[] J. Listgarten, M. Weinstein, B. P. Kleinstiver, A. A. Sousa, J. K. Joung, J. Crawford, K. Gao, L. Hoang, M. Elibol, J. G. Doench, and N. Fusi, *Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs,* Nature Biomedical Engineering **2**, 38 (2018).

[] M. Haeussler, K. Schönig, H. Eckert, A. Eschstruth, J. Mianné, J. B. Renaud, S. Schneider-Maunoury, A. Shkumatava, L. Teboul, J. Kent, J. S. Joly, and J. P. Concordet, *Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR,* Genome Biology **17**, 1 (2016).

I

Target recognition and off-target prediction

2

Hybridization kinetics explains CRISPR-Cas off-targeting rules

Due to their specificity, efficiency, and ease of programming, CRISPR associated nucleases are popular tools for genome editing. On the genomic scale, these nucleases still show considerable off-target activity though, posing a serious obstacle to the development of therapies. Off-targeting is often minimized by choosing especially high-specificity guide sequences, based on algorithms that codify empirically determined off-targeting rules. A lack of mechanistic understanding of these rules has so far necessitated their ad hoc implementation, likely contributing to the limited precision of present algorithms. To understand the targeting rules, we kinetically model the physics of guide-target hybrid formation. Using only four parameters, our model elucidates the kinetic origin of the experimentally observed off-targeting rules, thereby rationalizing the results from both binding and cleavage assays. We favorably compare our model to published data from CRISPR-Cas9, CRISPR-Cpf1, CRISPR-Cascade, as well as the human Argonaute 2 system.

This chapter has been published as: **M.Klein**, B.Eslami-Mossallam, D.Gonzalez Arroyo and M.Depken. Hybridization kinetics explains CRISPR-Cas off-targeting rules. *Cell Reports* **22** 1413–1423 (2018)

2.1. Introduction

RNA guided nucleases (RGNs) target nucleic-acid sequences based on complementarity to any guide RNA (gRNA) loaded into the complex. This versatility, together with the ability to design synthetic gRNA complementary to any target of choice, holds great promise for gene editing and gene silencing applications [**? ?**]. Among the known RGNs, the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) associated (Cas) nucleases Cas9 [**? ? ? ?**] and Cpf1 [**?**] are of special interest, as they are comparatively simple single-subunit enzymes.

Cas nucleases originate from the CRISPR-Cas adaptive immune system, which many prokaryotes use to fight off foreign genetic elements. *In vivo*, the Cas protein (complex) is programmed by loading RNA transcribed from a CRISPR locus in the host genome. The transcribed sequence includes sections referred to as spacers, which were acquired during past encounters with foreign genetic elements [**?**]. Once programmed, the Cas nuclease is able to target and degrade genetic elements with the same sequence as the stored spacer, and so offers protection against repeat invasions. An autoimmune response to sequences stored at the CRISPR locus is prevented through the additional requirement of a protein-mediated recognition of a short protospacer-adjacent motif (PAM) sequence present in the foreign genome, but not incorporated into the CRISPR locus with the spacer [**? ?**].

As viruses evolve in response to the selective pressure induced by the CRISPR-Cas immune system, the host is in turn under pressure to attack slightly mutated target sequences in addition to the target. It is therefore not surprising that Cas nucleases exhibit considerable off-target activity on sequences similar to the intended target [**? ? ? ? ? ? ? ? ? ?**]. Such off-targeting presents a severe problem for therapeutics, as DNA breaks introduced at the wrong site could lead to loss-of-function mutations in a well-functioning gene, or the improper repair of a disease causing gene [**?**].

To shed light on the determinants of off-target activity, a recent flurry of experiments has probed the level of binding and/or cleavage on mutated target sequences: high-throughput screens of large libraries of off-targets [**? ? ? ? ? ?**], genome-wide identification [**? ? ? ? ? ? ? ? ?**], systematic biochemical studies [**? ? ? ? ? ? ? ? ? ?**], structural studies [**? ? ? ? ? ? ?**], and single-molecule biophysical studies [**? ? ? ? ? ? ?**] providing insights into the mechanics of targeting. To date, a number of rather peculiar targeting rules have been empirically established for Cas nucleases: (i) seed region: single mismatches within a PAM proximal seed region can completely disrupt interference [**? ?**], while PAM distal mismatches have much less of an effect [**? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?**]; (ii) mismatch spread: when mismatches are outside the seed region, off-targets with spread out mismatches are targeted most strongly [**? ? ? ?**]; (iii) Differential binding vs. differential cleavage: binding is more tolerant to mismatches then cleavage [**? ? ? ? ? ? ?**]. (iv) specificity-efficiency decoupling: weakened protein-DNA interactions can improve target selectivity while still maintaining efficiency [**? ? ? ?**]. Although these experimental observations have already aided the development of strategies to improve the specificity of the CRISPR-Cas9 system [**? ? ? ? ?**], an understanding of the mechanistic origin behind target selectivity is still lacking, and our ability to predict off-targets remains limited [**? ? ? ?**].

Current off-target prediction algorithms are often based on sequence alignment with the target, and discard potential targets if they have more than some (user-defined) threshold number of mismatches [**? ? ? ?**]. To recover the mismatch-position dependence observed as seed regions (rule (i)) and their cooperativity (rule (ii)), such scoring schemes must be supplemented with ad hoc rules that penalize seed and closely spaced mismatches more than non-seed mismatches [**? ?**]. To move beyond ad hoc scoring schemes, we here use biophysical modelling to incorporate knowledge of the underlying targeting process. With this aim, it would be attractive to assume that the binding dynamics has had time to equilibrate before DNA degradation [**? ?**], as this would allow us to use simple binding/hybridization energetics to predict cleavage activity. Though attractive, this approach has recently been questioned by Bisaria *et al.* by noting that off-rates are generally not found to be much faster than cleavage rates [**?**], as would be required for establishing a binding equilibrium before cleavage. In addition, the authors show how abandoning the equilibration assumption directly explains the specificity increase observed with shortened gRNA [**?**].

Inspired by these observations, we go beyond binding energetics to build a biophysical model capturing the kinetics of guide-target hybrid formation. We show that the targeting rules (i)-(iv) can be seen as simple consequences of kinetics. The targeting rules are captured by four parameters that pertain to transition barriers between metastable states of the nuclease-guide-target complex, and we translate these into four experimentally observable quantities: the length of the seed region, the width of the transition region from seed to non-seed, the maximum amount of cleavage on single-mismatch off-targets, and the minimal distance between mismatches outside the seed region that allows for the cleavage of targets with multiple mismatches. By tying microscopic properties to biological and technological function we here open the door to refined and rational reengineering of the CRISPR-Cas system to further its use in therapeutic applications.

Though we frame our considerations in terms of the well-studied and technologically important Cas9, our approach applies to any RGN that displays a progressive matching between guide and target before cleavage (**Figure ??**A). To demonstrate the generality and power of our approach, we present fits to targeting data from Argonaute 2 (hAgo2), as well as type I, II and V CRISPR systems.

2.2. Results

At the start of target recognition, Cas nucleases bind to dsDNA from solution. The subsequent recognition of a PAM sequence triggers the DNA duplex to open up (**Figure ??**A), exposing the PAM proximal nucleotides to base pairing interactions with the guide [**? ?**]. From here, an R-loop is formed, expanding the guide-target hybrid in the PAM distal direction [**? ? ? ? ? ?**]. If the target and guide reach (near-) full pairing, cleavage of the two DNA strands is triggered [**?**].

To establish the determinants of off- vs. on-target cleavage, we construct a biophysical model of sequential target recognition in the unsaturated binding regime (see Methods). Using this model, we can calculate the rate of cleavage for off-targets, given the guide. To

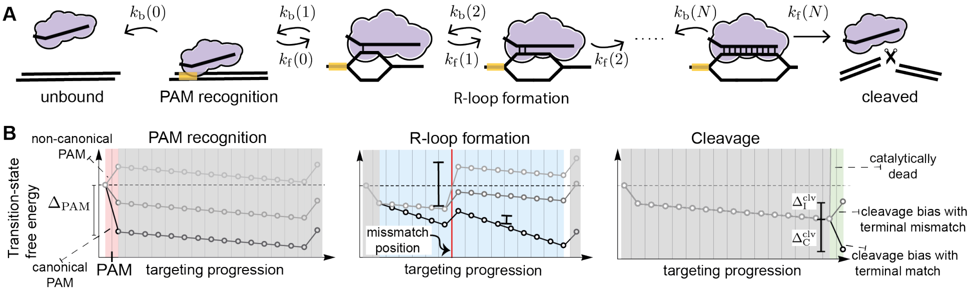


Figure 2.1: Kinetic model of RGN target recognition. **(A)** The RGN initially binds its' substrate at the PAM site, from which it can either unbind with rate $k_\mathrm{b}(0)$, or initiate R-loop formation with rate $k_\mathrm{f}(0)$. A partially formed R-loop of length n grows to length $n+1$ with rate $k_\mathrm{f}(n)$, or shrinks to length $n-1$ with rate $k_\mathrm{b}(n)$. Eventually, the RGN will either cleave its substrate with rate $k_\mathrm{f}(N)$ or reject the substrate and unbind with rate $k_\mathrm{b}(N)$. In the special case of a RGN that does not utilize PAM binding, it is assumed to bind straight into the initial state of R-loop formation. **(B)** The transition landscape of our minimal model. In the left panel, we illustrate a PAM bound enzyme kinetically biased toward R-loop formation by different amounts (black, grey, and light grey curves). The kinetic bias for the canonical PAM shown as Δ_PAM. In the middle panel we illustrate two kinetic biases toward R-loop extension (black and grey curves), with the larger bias indicated as Δ_C. In the same panel we further illustrate two kinetic biases against R-loop extension (grey and light grey curves) at mismatches (red vertical lines), with the largest bias shown as Δ_I. Once the complete R-loop is formed, the system is kinetically biased against cleavage by $\Delta^\mathrm{clv}_\mathrm{C/I} = \Delta_\mathrm{C/I} \mp \Delta_\mathrm{clv}$, as dictated by the nature of the terminal base pairing. See **Figure ??** for complete energy landscapes.

incorporate the mechanics of hybrid formation, we envision the changing extension of the R-loop as a diffusion through a free-energy landscape, eventually ending in either unbinding from, or degradation of, the targeted sequence (**Figure ??**A-B). Our model is parameterized by the free-energy of transition states surrounding the metastable states of PAM binding and the different progressions of R-loop formation (see Methods and section **??**). When in a metastable state, the RGN will be biased towards transitioning to the neighboring state with the lowest intervening barrier. The difference in heights of the surrounding barriers thus encodes the directions in which the system is most likely to progress, and we therefore refer to these differences as kinetic biases (**Figure ??**C). The balance between eventual unbinding or cleavage can be calculated with reference to kinetic biases alone, and visualized by a 'transition landscape' tracing out the transition states (**Figure ??**B, **??** and Methods). In such a landscape, the R-loop typically grows whenever the forward barrier is lower than the backward barrier; that is, whenever the transition landscape tilts downward. To facilitate the discussion of our exact results, we appropriate a rule-of-thumb from the limit of large biases (Methods): after binding the PAM, Cas9 is most likely to unbind before cleavage if the highest barrier to cleavage is greater than the highest barrier to unbinding, and vice versa (**Figure ??**A-B).

Though we treat the general scenario in the Methods section, we here further limit ourselves to a minimal description with only four effective microscopic parameters, pertaining to the average kinetic bias for: R-loop initiation after PAM binding (Δ_PAM), R-loop extension past a correctly matched (Δ_C) and mismatched (Δ_I) base pairs, and additional bias against cleavage once the R-loop is fully formed (Δ_clv) (for definitions see **Figure ??**B and Methods). The parameter Δ_clv is chosen such that the forward barrier after R-loop completion is independent of the nature of the terminal base (Methods), setting the final bias against cleavage to $\Delta_{\text{C/I}}^{\text{clv}} = \Delta_{\text{C/I}} \mp \Delta_{\text{clv}}$ (**Figure ??**B). Using this approach, we investigate to what extent our minimal model explains the four empirical targeting rules deduced from experiments.

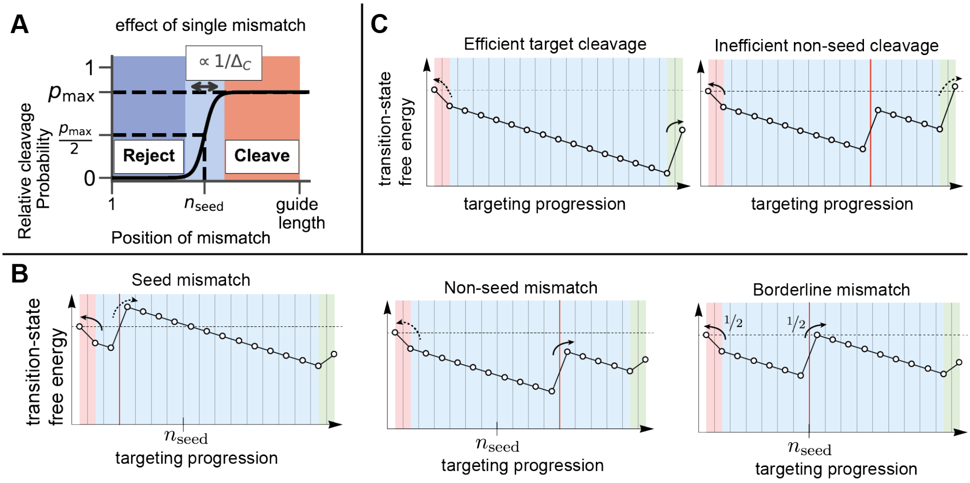


Figure 2.2: Rule (i) – seed region. **(A)** The relative-to-wildtype cleavage probability of a target with a single mismatch. Our model predicts a sigmoidal curve, with maximum off-target activity p_{max}, seed length n_{seed}, and width of the seed to non-seed transition $\sim 1/\Delta_{\text{C}}$. See figure **??** for parametric sweeps. **(B)** Transition landscapes illustrating that the placement of a single mismatch (fltr: before, exactly at, beyond the seed's border) influences the cleavage probability. **(C)** Increasing the kinetic bias against cleavage can suppress cleavage of off-targets with a PAM distal mismatch (compare right panel to right panel in (B)), while still maintaining a high on-target activity (left panel).

2.2.1. Rule (i): Seed region

Following PAM binding, base pairing between guide and target is attempted (**Figure ??**B; middle panel). To establish if the above mentioned dependence of the cleavage propensity on the position of mismatches within the guide-target hybrid could originate from the kinetics of the targeting process, we calculate the relative cleavage probability on a sequence with a single mismatch at position , compared to the cleavage probability on the target sequence. In section **??** we show that this relative cleavage probability is in general sigmoidal

$$p_{\text{clv}}(n) = \frac{p_{\text{max}}}{1 + \exp\left[-(n - n_{\text{seed}})\Delta_{\text{C}}\right]}, \tag{2.1}$$

with n_{seed} giving the position where the cleavage probability is half that of its maximum p_{max} (**Figure ??**A), and the biases are measured in units of $k_B T$. We identify n_{seed} as the length of the kinetic seed region, beyond which a mismatch will no longer strongly suppress cleavage (**Figure ??**A). From Equation **??** we see that the width of the transition from seed to non-seed region directly reports on the (average) correct-match bias (Δ_{C}, see section **??**), becoming narrower as the bias increases (**Figure ??**A and **Figure ??**A).

The emergence of a seed-like region can be understood from considering the rule-of-thumb that the fate of the enzyme is dictated by the largest barrier: when a mismatch is placed at n_{seed} (**Figure ??**B; right panel), the highest barrier to cleavage matches the barrier towards unbinding, guaranteeing a near equal probability for cleavage and unbinding. Placing the mismatch closer to the PAM increases the highest barrier towards cleavage (compare highest node to first node in **Figure ??**B; left panel), increasing the probability of rejecting such off-targets. Moving the mismatch distally from the PAM will gradually lower the highest barrier towards cleavage (**Figure ??**B; middle panel), increasing the probability of accepting such off-targets. Though the exact form of the parameters of Equation 1 are given in the Supplemental Information, it is informative to here give the kinetic seed length in the large-bias limit (Methods, **??**),

$$n_{\text{seed}} \approx \frac{\Delta_{\mathsf{I}} - \Delta_{\mathsf{PAM}}}{\Delta_{\mathsf{C}}} + 1 \tag{2.2}$$

From this we see that PAM bias and the base pairing biases all contribute to setting the extent of the seed region (**Figure ??**A, **??**B). Weakening the PAM or correct-match bias extends the seed region, while weakening the bias for incorrect matches shrinks it.

After PAM recognition and R-loop formation, cleavage completes a successful targeting process (**Figure ??**B; right panel). Tuning the final transition state allows us to toggle between different regimes of minimal single-mutation specificity. Targets with a PAM distal mismatch get cleaved with near unity probability ($p_{\mathsf{max}} \approx 1$) only if all transition states towards cleavage (including the cleavage step) lie well below the transition state to unbinding (**Figure ??**C; left panel, **Figure ??**C). For slow enough enzymatic activity, the final barrier towards cleavage might not go far below the barrier to unbinding, limiting the maximal cleavage compared to the perfect match ($p_{\mathsf{max}} < 1$)(**Figure ??**C; right panel). Consequently, there can be a noticeable effect on off-target activity also when the mismatch is outside the seed region (**Figure ??**A, **??**C). Reversing this logic implies that a $p_{\mathsf{max}} < 1$ is indicative of a relatively slow cleavage reaction.

2.2.2. Rule (ii): Mismatch spread

Considering more complex mismatch patterns, we start by addressing all possible dinucleotide mismatches (**Figure ??**A and **??**B). The overall cleavage and binding patterns obtained strongly resemble experimental observations [**? ? ?**]. As expected, placing both mismatches within the seed disrupts cleavage (**Figure ??**A). However, moving the mismatches outside the seed does not necessarily restore cleavage activity.With the first mismatch outside the seed region, a second mismatch only abolishes cleavage if it is situated before $n_{\text{seed}} + n_{\text{pair}}$ (**Figure ??**B), with

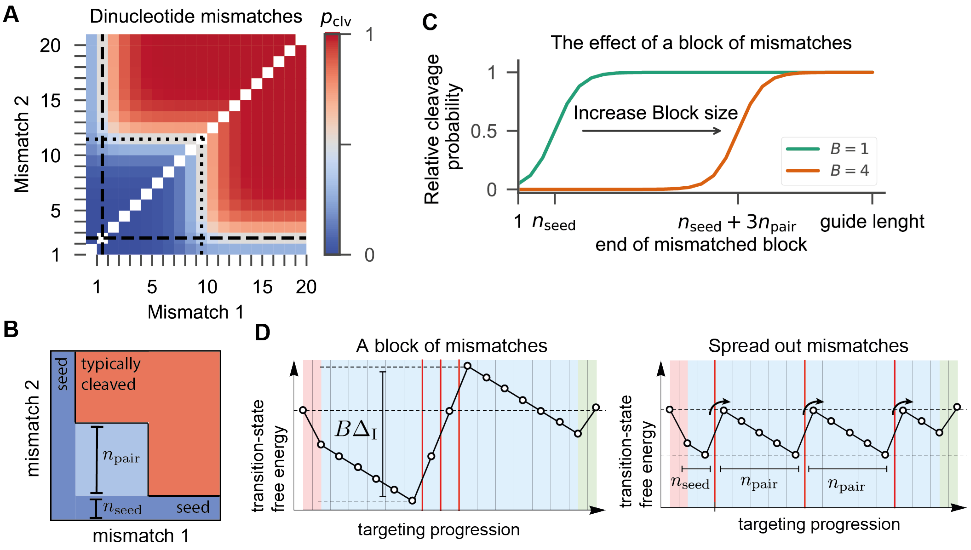


Figure 2.3: Rule (ii) – mismatch spread. **(A)** The relative-to-wildtype probability to cleave a target with two mismatches for a system with $\Delta_{\text{PAM}} = 3.5k_BT$,$\Delta_{\text{I}} = 4k_BT$,$\Delta_{\text{C}} = 1k_BT$ and $\Delta_{\text{clv}} = 1k_BT$. The seed length n_{seed} is indicated with dashed lines, and $n_{\text{seed}} + n_{\text{pair}}$ is indicated with dotted lines. **(B)** Schematic of the probability to cleave a target with two mismatches. The target is typically rejected in both blue regions and rejected in the red. **(C)** Probability to cleave a target with a block of B mismatches as a function of the location of the last mismatch. Also see **??**. **(D)** Spreading out blocked mismatches (left panel) around their average position significantly lessens the barrier to cleavage (right panel).

$$n_{\text{pair}} \approx \frac{\Delta_{\text{I}}}{\Delta_{\text{C}}} + 1, \tag{2.3}$$

in the large-bias limit (Methods and section **??**). The general form of the two-mismatch seed region is shown in Figure 3B, where only off-targets in the red region lead to cleavage. In the dark blue region, off-targets are rejected due to the first mismatch, and in the light blue region they are rejected due to the second mismatch. The single- and double-mismatch rules can now be unified and generalized (see **Figure ??**D; right panel) into a single rule for any number of mismatches: *"Off-targets will typically be rejected if any mismatch, say the mth mismatch, is positioned closer than $n_{\text{seed}} + (m-1)n_{\text{pair}}$ to the PAM."*. Note that for systems not requiring PAM recognition,$n_{\text{seed}} = n_{\text{pair}}$. The above rule also captures the extreme case of a 'block' of consecutive mismatches, which has also been investigated experimentally [**? ? ? ?**]. Placing such a block effectively acts as placing a single mismatch with the bias Δ_{I} scaled by the size of the block (**Figure ??**C-D and **Figure ??**), giving a block-seed region of size $n_{\text{seed}} + (B-1)n_{\text{pair}}$. Hence, a block of mismatches leads to less off-targeting compared to spread out mismatches (Figs 3C-D). Given the correspondence of these predictions with literature, our model seems to automatically and correctly capture the non-multiplicative cleavage suppression by multiple mismatches, in sharp contrast to the ad hoc scoring schemes employed in current prediction algorithms [**?**].

2.2.3. Rule (iii): Differential binding vs. differential cleavage

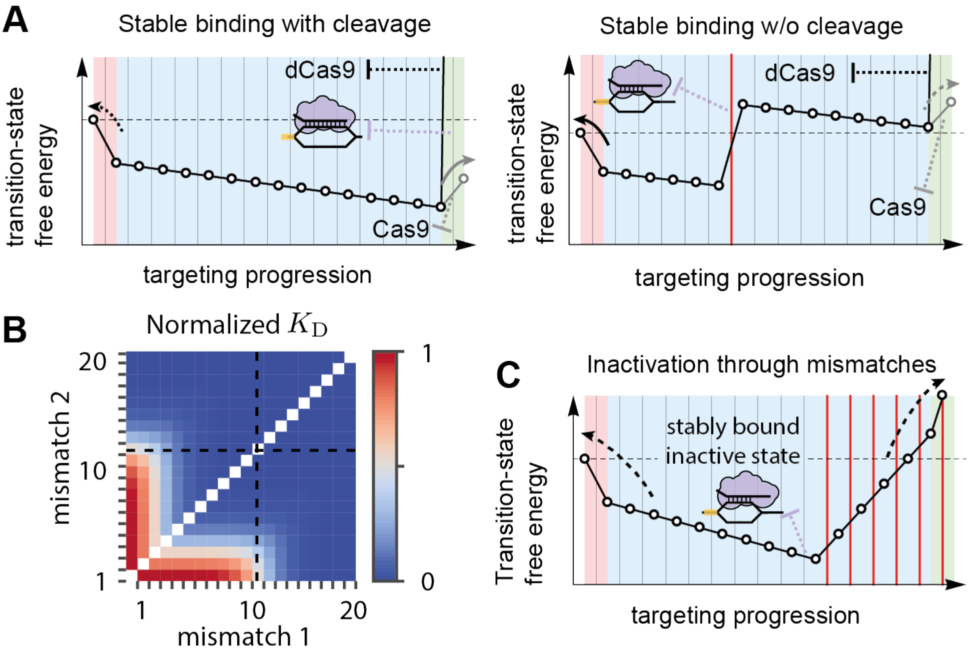


Figure 2.4: Rule (iii) - Differential binding versus differential cleavage. **(A)** Transition landscapes illustrating the difference between active Cas9 (grey curves) and dCas9 (black curves) when encountering either the cognate site (left panel) or an off-target with a mismatch within the seed (right panel). **(B)** The dissociation constant for targets with any combination of two mismatches for energetic biases $\delta_{\text{PAM}} = 7.5k_BT, \delta_{\text{C}} = 1k_BT$ and $\delta_{\text{I}} = 8k_BT$. The end of the seed region is indicated with dashed lines. See figure **??**for single-mismatched off-targets. **(C)** Transition landscape for an active Cas9 bound to an off-target possessing a block of mismatches placed at the PAM distal end. Even though cleavage is unlikely, unbinding takes a long time.

Catalytically dead systems (for example dCas9 [**?**] or Cascade without Cas3) bind strongly to sites that their catalytically active counterparts do not cleave [**? ? ? ? ? ?**]. In order to explain this effect, we model inactive systems with a very large cleavage barrier (gray in **Figure ??**B; right panel, Methods). In agreement with experimental observations [**?**], our model predicts a dissociation constant that is higher when a mismatch is placed closer to the PAM (**Figure ??**B and **??**).

In general, the gene editing (Cas9) and gene silencing (dCas9) capabilities should be seen as two related but separate properties of the RGN. For example, the most stable configuration of the RGN on the mismatched target shown in the right panel of **Figure ??**A is a bound state with a partial R-loop (purple). However, a catalytic active variant will most likely eventually reject this off-target (gray) as the barrier to cleavage is higher than to unbinding. Hence, even though cleavage sites are strong binders (**Figure ??**A; left panel), observing a long binding time on an off-target site should not be taken to imply that this site will also display substantial off-target cleavage (**Figure ??**A; right panel).

Active Cas9 variants also strongly bind to sites they are incapable of cleaving, especially those containing multiple PAM-distal mismatches [**? ?**]. Such a series of mismatches induces a large barrier that opposes, and thereby likely prevents, cleavage (**Figure ??**C). Although we are yet to extract temporal information from our model, it is clear that the state right before the first mismatch (purple) might be stably bound over experimental timescales.

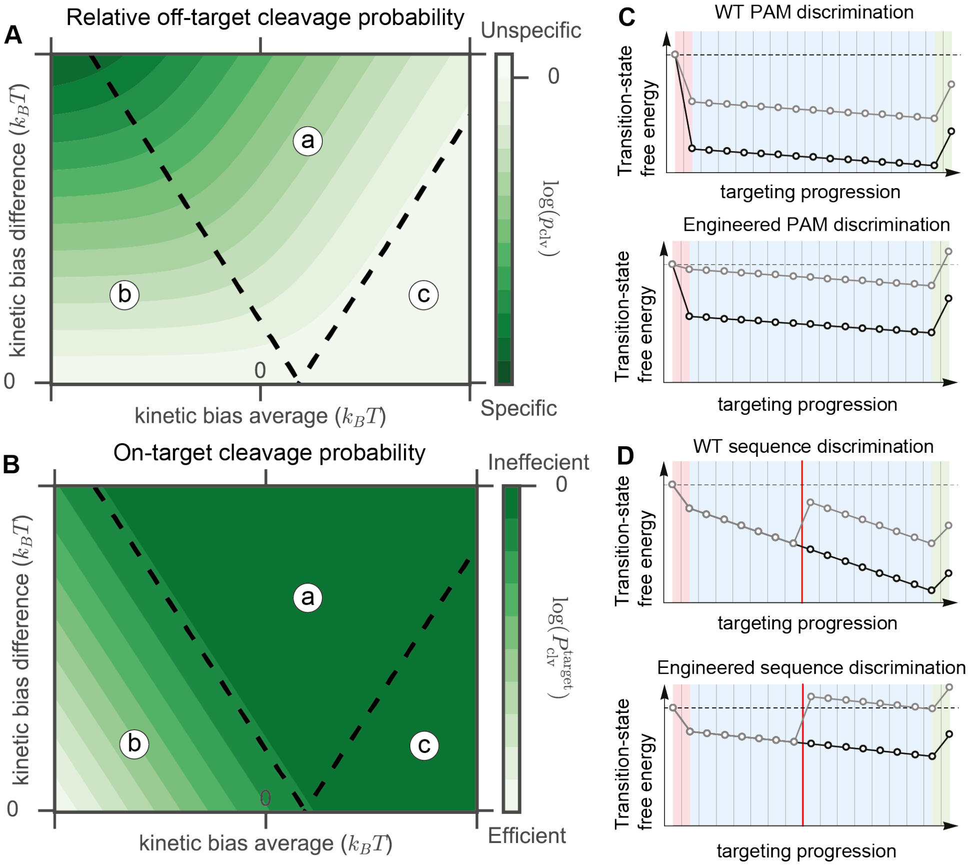


Figure 2.5: Rule (iv) – specificity-efficiency decoupling. (A) The cleavage probability on a fully cognate target but with a mismatched PAM, compared to one with the correct PAM, as a function of the average and difference in the kinetic bias of the correct and incorrect PAM. Independent of the sequence following both PAMs, one can identify three regimes (Supplemental Information). Only in regime a is the RGN's specificity improved through a decrease in the average PAM bias toward R-loop initiation. **(B)** On-target efficiency for the target with the correct PAM. In regime a, the RGN's efficiency is not compromised, allowing for simultaneous maintenance of on-target efficiency and specificity. **(C)** The cognate protospacer flanked by either a canonical PAM (black) or incorrect PAM sequence (grey) is bound by a WT (top panel) or engineered RGN (panel). **(D)** A matched/mismatched protospacer (black/grey) bound by wildtype/engineered RGN (top/bottom panel).

2.2.4. Rule (iv): Specificity-efficiency decoupling

R-loop formation is preceded by PAM recognition. Although PAM mismatches often completely abolish interactions with the target [? ? ?], binding to (and interference with)

targets flanked by non-canonical PAM sequences has been observed [**?**]. Since PAM mismatches will shift the entire free-energy landscape upwards from the bound PAM state onwards (**Figure ??**B; left panel), these always increase the highest barrier to cleavage, thereby reducing the cleavage efficiency on any sequence. For increased specificity, we thus need the cleavage efficiency for the off-targets to be reduced more than for the target itself.
Protein reengineering approaches most easily affect the overall strength of PAM interactions, influencing the kinetic bias for both the correct PAM (Δ_{PAM}) and incorrect PAM (Δ'_{PAM})). In **Figure ??**A we show the relative cleavage efficiency between protospacers flaked by incorrect and correct PAMs, and in **Figure ??**B we show the cleavage efficiency with the correct PAM — both as functions of the average kinetic bias ($(\Delta_{\text{PAM}} + \Delta'_{\text{PAM}})/2$) and the kinetic bias difference ($\Delta_{\text{PAM}} - \Delta'_{\text{PAM}}$). As long as the system operates in region A (**Figure ??**A), it is possible to increase the specificity by lowering the average kinetic bias toward R-loop formation without changing the kinetic-bias difference (section **??**). Outside this region, the system either does not discriminate between PAMs (region C) or is insensitive to the average kinetic bias (region B). Interestingly, it is only in region B that lowering the average bias also leads to a lower on-target efficiency (**Figure ??**B), and consequently the wild type (wt) nuclease can only be improved if brought into region A, where it is possible to engineer specificity increases with limited costs in the on-target efficiency. The transition-state diagrams shown in the top panel of **Figure ??**C show a situation where the barrier to cleavage (right most node) is substantially lower than the barrier to unbinding (leftmost node) for two different PAM biases, both resulting in near unit-probability to cleave , and corresponding to region C in **Figure ??**A. Reengineering the nuclease to have overall weaker PAM binding (**Figure ??**C, bottom panel) brings the system into region B, where the cleavage probability for the correct PAM (black) remains close to unity, while the probability of cleaving with the incorrect PAM (gray) is drastically lowered. The above scenario might explain how PAM mutant Cas9s are able to outperform their wildtype counterparts [**? ?**] on specificity without significant loss in efficiency.

Another approach to gain specificity is to weaken the protein-DNA interactions effecting the bias for R-loop extension [**? ?**]. In Figure 5D we show how engineering the PAM-bound nuclease in this way, inducing a lower gain for correct base pairing, can render previously cleaved off-targets (gray line in top panel) rejected (gray line in bottom panel). We further see how we can retain on-target specificity if the highest transition state towards cleavage (rightmost node of black line) remains substantially lower than the transition state to unbinding (leftmost node of black line). The above scenario might explain how mutant Cas9s could have an extended seed, while having negligible reduction in on-target cleavage activity [**? ?**].

2.2.5. Comparison to experimental data for a broad class of RNA guided nucleases

To test our model, we acquired published datasets from different RGN systems, and fitted Equation **??** to singly mismatched targets and blocks of mismatches. The fitted sigmoid has only three effective fit parameters (p_{max} or $K_{\mathrm{D,max}}$, n_{seed} and Δ_{C}), so we can unfortunately not get an estimate for all microscopic parameters from the single-mismatch datasets (sec-

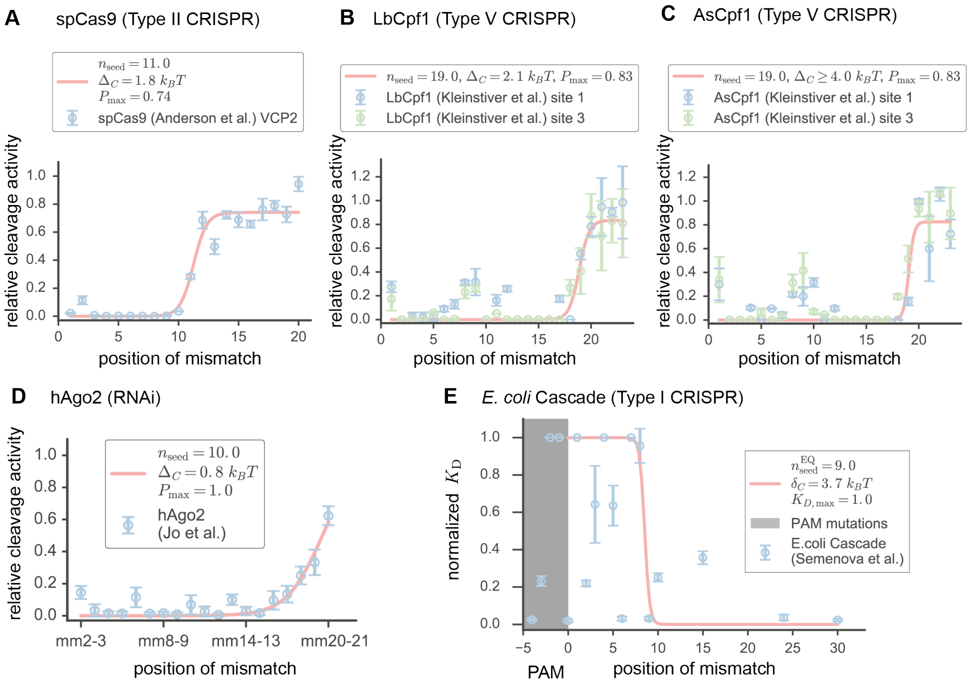


Figure 2.6: Comparison to experimental data. Fit of sigmoid (equation ??) to experimental data from: (A) spCas9 [?]. **(B)** LbCpf1 [?]. **(C)** AsCpf1 [?]. **(D)** Human Argonaute 2 [?]. **(E)** E. coli Cascade complex [?]. Values reported in (A)-(D) correspond to the median of 1000 bootstrap replicates, and the confidence intervals in the text correspond to 68%. See Figure **??** for additional fits.

tions **??** and **??**)—for this, further experiments are required, as outlined below. Details of the fitting procedure and additional fits can be found in section **??**.

Perhaps the best characterized RGN system is the Type-II CRISPR associated *Streptococcus Pyogenes* Cas9 (spCas9). Among the systems we estimate parameters for, the dataset from Anderson *et al.* [?] traces out the sigmoidal trend particularly well. For this data set we fit out a kinetic seed of about 11.3 [11.0,11.4] nt (68% confidence interval between 11.0 and 11.4), and an average bias per correct base pair of about $\Delta_{\rm C} = 1.70[1.15, 4.0]k_BT$ (**Figure ??**A). This positive bias indicates that association with the RGN stabilizes the hybrid, which is in line with recent studies demonstrating that the protein has a strong contribution to the energetics of the resulting bound complex [**? ? ?**]. The relative cleavage probability levels-off around $p_{\rm max} = 0.74[0.72, 0.77]$, indicating that spCas9 retains some specificity even against errors that are outside the seed. We performed additional fits using a second target site from the dataset of Anderson *et al.* and data obtained from Pattanayak *et al.* [**?**], which produced results that do not significantly differ (**Figures** S5A-C).

Recently, the type V CRISPR associated enzyme Cpf1 has been characterized as another single-subunit RGN [**?**]. Kleinstiver *et al.* [**?**] performed *in vivo* (human cells) cleavage assays using two different variants named LbCpf1 (**Figure ??**B) and AsCpf1 (**Figure ??**C). Both variants exhibit quantitatively similar off-targeting, both with seed lengths ($n_{\text{seed}} \approx 18.9[18.5, 19.2]$ nt for LbCpf1 vs. 19.1 [18.7,19.3] nt for AsCpf1) and maximum off-target activity ($p_{\text{max}} \approx 0.84[0.66, 1.0]$ nt for LbCpf1 vs. 0.83[0.71,1.0] for AsCpf1). Compared to spCas9, the Cpf1s are much more specific as the seed region is significantly larger.

Single-molecule FRET experiments done with hAgo2 [**?**] utilized targets with two consecutive mismatches. Given that hybrid formation is not preceded by PAM recognition, and that consecutive mismatches impose a combined penalty (**Figures** 3C-D), the estimated half-saturation point is approximately twice the kinetic seed length for a single mismatch ($n_{\text{seed}} \approx 10$ [9.5,9.9] nt). The hAgo2 data thus suggests a similar seed length as that of spCas9 (**Figure ??**D), consistent with the observation that hAgo2 and spCas9 display structural similarities within their respective seed regions [**?**]. Our fits further reveal that hAgo2 likely exhibits a substantially lower gain per correctly formed base pair ($\Delta_{\text{C}} \approx 0.77[0.66, 0.92]k_B T$).

Unlike the aforementioned RGNs, the Type I CRISPR uses a multi-subunit protein complex, termed Cascade, to target invaders [**?**]. Semenova *et al.* [**?**] measured the dissociation constant *in vitro* of the *E. Coli* subtype I-E Cascade. Fitting their data, we find that mismatches within the first 9 nt of the guide lead to rapid rejection (**Figure ??**E). Interestingly, the energetic gain for a match again suggests a large contribution of the protein to the overall stability (energetic bias $\delta_{\text{C}} \approx 3.7k_B T$). Structurally, subunits of the Cascade complex bind to nucleotides 6, 12, 18, 24 and 30 of the guide [**?**]. To model this property we assume that incorporating matches or mismatches at the Cascade-guide binding positions does not affect affinity. Including this effect mainly reduced the estimated energetic gain for matches ($\delta_{\text{C}} \approx 1.9k_B T$, section **??** and **Figure ??**D), a value more in line to those obtained for the other CRISPR systems.

2.3. Discussion

We have presented a general description of target recognition by RGNs with a progressive matching between guide and target (**Figure ??**A), applicable to both CRISPR and Argonaute systems. In its simplest form, our model contains only two parameters to describe the R-loop formation process: an average kinetic bias towards incorporation beyond a match (Δ_{C}) and an average kinetic bias against extending the R-loop beyond a mismatch (Δ_{I}) (**Figure ??**B; middle panel). Despite the simplifications going into this minimal model, we can qualitatively understand the targeting rules for these RGNs as resulting from kinetics, as illustrated graphically for: seed region (**Figure ??**B), mismatch spread (**Figure ??**D), the poor match between cleavage propensity and binding propensity (**Figure ??**A) and the specificity-efficiency decoupling (**Figure ??**C-D). Based on our model we have been able to establish a general targeting rule: *"Off-targets will typically be rejected if any mismatch, say the m^th mismatch, is positioned closer than* $n_{\text{seed}} + (m-1)n_{\text{pair}}$ *to the PAM."*

Although Figure 6 shows that our model can already describe experimental data from various RGNs, the number of microscopic parameters in the physical model (Δ_{PAM},Δ_{C} ,Δ_{I} and Δ_{clv}, Figure 1B) exceeds the number of fit parameters available from single-mismatch experiments (Δ_C,p_max, and n_seed). It is therefore not possible to determine all the microscopic parameters from single-mismatch experiments alone. However, Figure 3B shows that with two mismatches, we could also fit out n_pair, and so determine all the microscopic parameters. It should be possible to directly extract all four microscopic parameters once such extended datasets become available.

One should recognize that our minimal model does not capture all the physics of the targeting process. Nucleic-acid interactions are explicitly sequence dependent, RGNs are known to undergo conformational changes prior to cleavage [**? ? ?**], and the Δ_C we fit out in Figure 6 technically only reports the matching-bias at the end of the seed, allowing for variable biases along the R-loop. Although these are all topics that need to be explored for future improved quantitative predictions, such extensions are not needed to explain the observed targeting rules, and will not qualitatively alter the trends predicted by our model. An exception might be the data from Cpf1 (**Figure ??**B-C), since it shows an increased tolerance to mismatches of nucleotides 1,2,8 and 9 compared to our minimal model, with a second independent study showing the same behavior [**?**]. Similarly, deviations from the sigmoidal trend are observed for Cascade (**Figure ??**E). Such features could be explained either through a sequence or position dependence of the kinetic biases.

In conclusion, our model is capable of explaining the observed off-targeting rules of CRISPR and Argonaute systems in simple kinetic terms. After having established the general utility of this approach, the next step will be to move beyond our minimal model and gradually allow for conformational control and sequence effects by letting our parameters depend on the nature of matches/mismatches as well as their positions. Fitting such a generalized model against training data would likely improve on present target prediction algorithms by limiting overfitting, as it captures the basic targeting rules deduced from experiments while using only a minimal set of physically meaningful parameters.

2.4. Methods

2.4.1. A general model for RGNs with progressive R-loop formation followed by cleavage

Given the observed dependence of cleavage activity on Cas9 concentration [**? ? ? ? ?**], we here limit ourselves to the regime where nuclease concentrations are low enough that all binding sites are unsaturated. The unsaturated regime is also the regime with the highest specificity, and should therefore be of particular interest in gene-editing applications.

We define the cleavage efficiency $P_\mathrm{clv}(s|g)$ as the fraction of binding events to sequence s that result in cleavage, given the RGN is loaded with guide sequence g. If we in the unsaturated regime assume the binding rate to be independent of sequence, we can express the relative rate of non-target vs. target cleavage as

$$p_\mathrm{clv}(s|g) = \frac{P_\mathrm{clv}(s|g)}{P_\mathrm{clv}(g|g)} \tag{2.4}$$

This relative efficiency is a direct measure of specificity, approaching unity for non-specific targeting ($P_\mathrm{clv}(s|g) \approx P_\mathrm{clv}(g|g)$) and zero for specific targeting ($P_\mathrm{clv}(s|g) \ll P_\mathrm{clv}(g|g)$).

In our model, we denote the PAM bound state as and the subsequent R-loop states by the number of base pairs that are formed in the hybrid. Each of the states $n = 1, ..., N$ are taken to transition to state $n - 1/n + 1$ with backward/forward hopping rate $k_{\mathrm{b}}(n)/k_{\mathrm{f}}(n)$ (**Figure ??**A). The ratio between forward and backward rates sets the relative probability of going forward and backward from any state, and can be parametrized in terms of $\Delta(n)$, the difference in the free-energy barrier between going backwards and forwards from state n (**Figure ??**A),

$$\frac{k_{\mathrm{f}}(n)}{k_{\mathrm{b}}(n)} = e^{\Delta(n)}. \tag{2.5}$$

Here we measure energy in units of $k_B T$ for notational convenience, and we will refer to $\Delta(n)$ as the bias toward cleavage. The model (**Figure ??**A) is known as a birth-death process [**?**], and the cleavage efficiency is given by the expression (section **??**),

$$P_{\mathrm{clv}}(s|g) = \frac{1}{1 + \sum_{n=0}^{N} e^{-\Delta T(n)}}, \; \Delta T(n) = \sum_{m=0}^{n} \Delta(m). \tag{2.6}$$

Here $\Delta T(n)$ represents the free-energy difference between the transition-state to solution and the forward transition state from position n (**Figure ??**A-C). For systems like hAgo2, there is no initial PAM binding [**? ?**], and the sums in Equation **??** should omit the PAM state ($n, m = 0$).

2.4.2. Building intuition by using the transition landscape (large bias limit)

Though we will use the exact results of Equation **??** for all calculations, it is useful to build intuition for the system by considering the case of large biases. In this limit, the term (say $n = n^*$) with the highest transition-state dominates the sum in Equations **??** and **??** (**Figure ??**A-B), and the cleavage efficiency can be approximated as

$$P_{\mathrm{clv}}(s|g) \approx \frac{1}{1 + e^{-\Delta T(n^*)}} \tag{2.7}$$

Based on this we deduce the rule-of-thumb that cleavage dominates ($P_{\mathrm{clv}} > 1/2$) if the first state of the transition landscape is the highest ($\Delta T(n^*) > 0$) (**Figure ??**A). Conversely, a potential target is likely rejected ($P_{\mathrm{clv}} < 1/2$) if any of the other transition states lies above the first ($\Delta T(n^*) < 0$) (**Figure ??**B).

2.4.3. A minimal model for RGNs with progressive R-loop formation followed by cleavage

Given that the defining feature of RGNs is their ability to target any sequence, we expect the major targeting mechanisms to depend more strongly on mismatch position than on the precise nature of the mismatches. With this in mind, we consider a sequence independent model with the aim of finding a description that captures the gross, sequence averaged, features with a minimal number of parameters.

Focusing first on how PAM binding effects the system (**Figure ??**1; left panel), we see that

$\Delta(0) = \Delta_{\text{PAM}}$ controls the kinetic bias between initiating R-loop formation and unbinding. A canonical PAM (black) promotes R-loop initiation, while a non-canonical PAM lessens (darker gray) or reverses (lighter gray) the bias towards R-loop formation. Note that PAM independent systems omit this initial step.

Turning to the bias of R-loop progression, we represent the guide-target hybrid as a sequence of matches (C, correct base pairing) and mismatches (I, incorrect base pairing). Defining the average kinetic bias towards/against extending the R-loop by one correct/incorrect base pair as $\Delta_{\text{C}}/\Delta_{\text{I}}$ (**Figure ??**B; middle panel), we take $\Delta(n) = \Delta_{\text{C}}$ or $\Delta(n) = -\Delta_{\text{I}}$ depending on if the base pairing is correct or incorrect (section **??**). In the middle panel of Figure 1B we show a transition landscape with moderate gains for correct base pairings and moderate costs for incorrect base pairings (dark gray). The black transition landscape corresponds to an increased gain for matches, while the light gray corresponds to an increased penalty for mismatches.

Lastly, considering the bias between cleavage and unwinding of the R-loop, we assume that an incorrect base-pair at the terminal position adds the same change in bias as it did in the interior of the R-loop. Therefore, introducing the cleavage bias Δ_{clv}, we take $\Delta(N) = \Delta_{\text{C}}^{\text{clv}}$ for a correct match and $\Delta(N) = -\Delta_{\text{I}}^{\text{clv}}$ for a mismatch, with $\Delta_{\text{C/I}}^{\text{clv}} = \Delta_{\text{C/I}} \mp \Delta_{\text{clv}}$ as bias against cleavage from the fully hybridized state (**Figure ??**B; right panel). In the right panel of **Figure ??**B, we show examples where the terminal bias $\Delta_{\text{C/I}}^{\text{clv}}$ corresponds to a terminal match (black), terminal mismatch (dark gray), and for a catalytically dead nuclease (light gray).

2.4.4. Dissociation constant for catalytically dead nucleases

Apart from examining cleavage propensity, many experiments have focused on the binding of catalytically dead Cas9 (dCas9) or other catalytically dead RGNs [**? ? ? ? ? ? ?**]. To be able to relate pure binding experiments to cleavage experiments, we also calculate the dissociation constant K_{D} for our minimal model when describing a catalytically dead system ($\Delta_{\text{clv}} \approx \infty$) (**Figure ??**D) through

$$P_{\text{bound}} = \frac{[\text{RGN}]}{[\text{RGN}] + K_{\text{D}}} \tag{2.8}$$

Here P_{bound} equals the probability to bind a substrate in any of the (N) possible R-loop configurations and follows from Equation **??** (see section **??**). Further, $[\text{RGN}]$ denotes the concentration of effector complex. Differences in stability of the bound states now parameterize our model (Fig S1D).

2.5. Author Contributions

M.K and M.D designed the research. M.K, B.E.M and D.G.A performed the research and, together with M.D, interpreted the data. M.K, M.D and B.E.M wrote the manuscript. The authors declare no competing interests.

2.6. Acknowlegdements

First and foremost, MD would like to thank Ralph Seidel for introducing him to the problem of sequential target recognition during a visit to Delft. We would also like to thank

Aafke van den Berg, Michiel Bongaerts, and Kristian Blom for fruitful discussion regarding the theoretical modeling. We further acknowledge discussions we have had with all members of the CRISPR/microRNA community in Delft: Viktorija Globytė, Luuk Loeff, Tao Ju (Thijs) Cui, Stanley Chandradoss, Sungchul Kim, Seung Hwan Lee, Chirlmin Joo, Rebecca McKenzie, Jochem Vink, Sebastian Kieper and Stan Brouns. MK also thanks Iason Katechis for introducing him to the Python fitting library (lmfit), and Rebecca, Thijs and Chirlmin for carefully reading and commenting on the manuscript. We gratefully thank Emily Anderson, Benjamin Kleinstiver and Keith Joung, Sungchul Hohng, Soochul Shin and Myung Hyun Jo, Ekaterina Semenova as well as Vikram Pattanayak for sharing their data and answering all our questions. This work was supported by the Netherlands Organization for Scientific Research (NWO/OCW), as part of the Frontiers in Nanoscience program. MD acknowledges financial support from a TU Delft startup grant. This work (B.E.M.) forms part of the research programme 'Crowd management: The physics of genome processing in complex environments', which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

2.7. Supplemental Information

2.7.1. A general kinetic model for target recognition

In Figure **??**A we illustrate the states of our model. The RGN is described as either being unbound, bound to the PAM (in case of CRISPR systems), having formed an R-loop of length $n = 1, \dots, N$ or having cleaved its target substrate. Let us label these states as $i \in [-1, N+1]$, with N being the total length of the guide (target) sequence. Each state $i \in [0, N]$ has rates $k_{\mathrm{f}}(i)$ and $k_{\mathrm{b}}(i)$ associated with it for transitioning to $i+1$ and $i-1$ respectively.

The cleavage probability

The probability to cleave a target site once the substrate is bound (P_{clv}) is equivalent to the fixation probability of a Birth-Death process with absorbing states being the unbound and post-cleavage states [**?**]. As the derivation is fairly straight forward, we give it here for completeness. When starting with an R-loop of length $n-1$, we calculate the probability to cleave $P_{\mathrm{clv},n-1}$ before reducing the R-loop to a length of $n-2$. Counting all paths that take you from $n-1$ to $N+1$ we can construct a recursion relation for $P_{\mathrm{clv},n}$,

$$\begin{aligned} P_{\mathrm{clv},n} &= \sum_{m=0}^{\infty} \left(\frac{k_{\mathrm{f}}(n)}{k_{\mathrm{b}}(n-1) + k_{\mathrm{f}}(n)} (1 - P_{\mathrm{clv},n+1}) \right)^m \frac{k_{\mathrm{f}}(n)}{k_{\mathrm{b}}(n) + k_{\mathrm{f}}(n)} P_{\mathrm{clv},n+1} \\ &= \frac{P_{\mathrm{clv},n+1}}{\gamma_n + P_{\mathrm{clv},n+1}}, \quad \gamma_n = \frac{k_{\mathrm{b}}(n)}{k_{\mathrm{f}}(n)} \end{aligned},$$

or equivalently

$$\frac{1}{P_{\mathrm{clv},n}} = 1 + \frac{\gamma_n}{P_{\mathrm{clv},n+1}}. \tag{S2.1}$$

The boundary probability $P_{\mathrm{clv},N}$, representing the probability to cleave staring with a full R-loop and without reducing the R-loop's length, is given by a simple splitting probability

$$P_{\mathrm{clv},N} = \frac{k_{\mathrm{f}}(N)}{k_{\mathrm{f}}(N) + k_{\mathrm{b}}(N)} = \frac{1}{1 + \gamma_N}. \tag{S2.2}$$

Using equations **??** and **??** we have

$$\frac{1}{P_{\mathrm{clv},0}} = 1+\gamma_0 \frac{1}{P_{\mathrm{clv},1}} = 1+\gamma_0+\gamma_0\gamma_1 \frac{1}{P_{\mathrm{clv},2}} = 1+\gamma_0+\gamma_0\gamma_1+\gamma_0\gamma_1\gamma_2 \frac{1}{P_{\mathrm{clv},3}} = \ldots = 1+\sum_{n=0}^{N}\prod_{i=0}^{n}\gamma_i,$$

from which it follows that

$$P_{\mathrm{clv}} \equiv P_{\mathrm{clv},0} = \frac{1}{1+\sum\limits_{n=0}^{N}\prod\limits_{i=0}^{n}\gamma_i}. \tag{S2.3}$$

The transition landscape

We assign a free-energy F_i to each metastable state $i \in [0, N]$, and the transition state energy T_i to the highest free energy point on the reaction path from i to $i+1$, for $i \in [-1, N]$. Introducing the attempt rate k_0 we write the associated forward and backward rates as follows (all energies are measured in units of the thermal energy)

$$k_{\mathrm{f}}(i) = k_0 \exp(-(T_i - F_i)), \quad k_{\mathrm{b}}(i) = k_0 \exp(-(T_{i-1} - F_i)) \Rightarrow \gamma_i = \exp(-\Delta_i), \quad \Delta_i = T_{i-1} - T_i. \tag{S2.4}$$

In terms of transition-state free energies we can write **??** as

$$P_{\mathrm{clv}} = \frac{1}{1+\sum_{n=0}^{N}\exp(-\sum_{i=0}^{n}\Delta_i)} \equiv \frac{1}{1+\sum_{n=0}^{N}\exp(-\Delta T_n)}, \quad \Delta T_n = \sum_{i=0}^{n}\Delta_i. \tag{S2.5}$$

From the above it is clear that the cleavage probability depends only on the transition state energies, and not on the free energies of the metastable states. If we assume there to be one dominant minimal bias, say for $n = n^*$, then this can be approximated as

$$P_{\mathrm{clv}} \approx \frac{1}{1+\exp(-\Delta T_{n^*})}. \tag{S2.6}$$

which we will refer to as the large-bias limit.

2.7.2. A minimal kinetic model for target recognition

To understand what constitutes the targeting principles of RGNs, we introduce a simplified model where: for the PAM state ($i = 0$) we have $\Delta_0 = \Delta_{\mathrm{PAM}}$; for a partial R-loop ($i \in [1, N-1]$) we have $\Delta_i = \Delta_{\mathrm{C}}$ if the i:th base in the R-loop is correctly matched, and $\Delta_i = -\Delta_{\mathrm{I}}$ if mismatched; for a completed R-loop ($i = N$) we have $\Delta_N = \Delta_{\mathrm{C}} - \Delta_{\mathrm{clv}}$ if the terminal base is mismatched, and $\Delta_N = -\Delta_{\mathrm{I}} - \Delta_{\mathrm{clv}}$ if mismatched. An R-loop in which n base pairs are incorporated, out of which $n_{\mathrm{C}}(n)$ are forming correct Watson-Crick pairs, is then described by

$$\Delta T_n = \Delta_{\mathrm{PAM}} + n_{\mathrm{C}}(n)\Delta_{\mathrm{C}} - (n - n_{\mathrm{C}}(n))\Delta_{\mathrm{I}} - \delta_{n,N}\Delta_{\mathrm{clv}}, \quad n = 0, \ldots, N \tag{S2.7}$$

where $\delta_{n,N}$ represents the Kronecker delta: $\delta_{n,N} = 1$ if $n = N$ and $\delta_{n,N} = 0$ otherwise. For PAM independent systems, we instead use

$$\Delta T_n = n_{\mathrm{C}}(n)\Delta_{\mathrm{C}} - (n - n_{\mathrm{C}}(n))\Delta_{\mathrm{I}} - \delta_{n,N}\Delta_{\mathrm{clv}}, \quad n = 1, \ldots, N.$$

The emergence of a seed region

Here we show that when comparing off-targets with a single mismatch to the cognate sequence, the relative cleavage probability is sigmoidal, irrespective of the values of the model parameters. Let there be a single mismatch at position n_{MM}, giving

$$n_{\mathrm{C}}(n) = \begin{cases} 0, & n < n_{\mathrm{MM}} \\ 1, & n \geq n_{\mathrm{MM}} \end{cases} .$$

Using equation **??** it is then straight forward to show that

$$p_{\mathsf{clv}}(n_{\mathrm{MM}}) \equiv \frac{P_{\mathsf{clv}}(\text{single error at } n_{\mathrm{MM}})}{P_{\mathsf{clv}}(\text{no error})} = \frac{p_{\mathsf{max}}}{1 + e^{-\Delta_{\mathsf{C}}(n_{\mathrm{MM}} - n_{\mathrm{seed}})}}, \tag{S2.8}$$

where

$$\begin{aligned} p_{\mathsf{max}} &= \frac{(1 - e^{-\Delta_{\mathsf{C}}})e^{\Delta_{\mathsf{PAM}}}(1 + e^{-\Delta T_N^{\mathrm{on}}}) + 1 - e^{-\Delta R_N^{\mathrm{on}}}}{(1 - e^{-\Delta_{\mathsf{C}}})e^{\Delta_{\mathsf{PAM}}}(1 + e^{-\Delta T_N^{\mathrm{tm}}}) + 1 - e^{-\Delta R_N^{\mathrm{tm}}}} \\ n_{\mathrm{seed}} &= \frac{1}{\Delta_{\mathsf{C}}} \ln \left[\frac{e^{\Delta_{\mathsf{I}} + \Delta_{\mathsf{C}}} - 1}{(1 - e^{-\Delta_{\mathsf{C}}})e^{\Delta_{\mathsf{PAM}}}(1 + e^{-\Delta T_N^{\mathrm{tm}}}) + 1 - e^{-\Delta R_N^{\mathrm{tm}}}} \right], \end{aligned} \tag{S2.9}$$

and we have introduced the R-loop completion bias with a cognate and terminal-mismatch target respectively

$$\Delta R_N^{\mathrm{on}} = N\Delta_{\mathsf{C}}, \quad \Delta R_N^{\mathrm{tm}} = (N-1)\Delta_{\mathsf{C}} - \Delta_{\mathsf{I}} = \Delta R_N^{\mathrm{on}} - (\Delta_{\mathsf{C}} + \Delta_{\mathsf{I}})$$

as well as the total bias toward cleavage of the on-target and on off-target with terminal-mismatch target respectively

$$\Delta T_N^{\mathrm{on}} = \Delta R_N^{\mathrm{on}} + \Delta_{\mathsf{PAM}} - \Delta_{\mathsf{clv}}, \quad \Delta T_N^{\mathrm{tm}} = \Delta R_N^{\mathrm{tm}} + \Delta_{\mathsf{PAM}} - \Delta_{\mathsf{clv}} = \Delta T_N^{\mathrm{on}} - (\Delta_{\mathsf{C}} + \Delta_{\mathsf{I}}).$$

Here p_{max} represents an upper bound on the achievable relative cleavage rate, and n_{seed} marks the transition from a region with no cleavage (the seed region) to a region with maximal cleavage. Note that our sigmoid function has three parameters (p_{max}, n_{seed} and Δ_{C}), which is one less than then number of microscopic parameters (Δ_{PAM}, Δ_{C}, Δ_{I}, and Δ_{clv}). Hence, we will not be able to fit out all four microscopic parameters relying on single-mismatch-data alone. Interestingly, the microscopic parameter Δ_{C} also sets the with of the transition region from seed to non-seed. To get an estimate of the width of the transition region, we linearize p_{clv} around the point of most rapid increase ($n_{\mathrm{MM}} = n_{\mathrm{seed}}$)

$$p_{\mathsf{clv}}(n_{\mathrm{MM}}) \approx \frac{1}{2}p_{\mathsf{max}} + \frac{1}{4}p_{\mathsf{max}}\Delta_{\mathsf{C}}(n_{\mathrm{MM}} - n_{\mathrm{seed}}). \tag{S2.10}$$

This function transitions from no relative cleavage to maximal relative cleavage over the distance $w = 4/\Delta_{\mathsf{C}}$, giving us an estimate of the width of the transition region.

When dealing with a stretch of mismatches, the relative cleavage probability still follows the sigmoidal form of equation **??**, but with modified p_{max} and n_{seed}.

The physiological limit and the large-bias limit

For the correct PAM we expect there to be a considerable PAM bias, and assuming at least a moderate bias for R-loop extension over correct basepairs, we should be able to take $(1-e^{-\Delta_{\mathsf{C}}})e^{\Delta_{\mathsf{PAM}}} \gg 1$ in equation **??**. Further, we expect the overall bias on an on-target to be strongly toward cleavage ($\Delta T_N^{\mathrm{on}} \gg 1$), as well as a large change in total bias when comparing a correctly and incorrectly matched base pair ($\Delta_{\mathsf{I}} + \Delta_{\mathsf{C}} \gg 1$). With these assumptions equation **??** becomes

$$\begin{aligned} p_{\mathsf{max}} &\approx \frac{1}{1+e^{-\Delta T_N^{\mathrm{tm}}}} \\ n_{\mathrm{seed}} &\approx \frac{\Delta_{\mathsf{I}}+\Delta_{\mathsf{C}}-\Delta_{\mathsf{PAM}}}{\Delta_{\mathsf{C}}} + \frac{\ln p_{\mathsf{max}} - \ln(1-e^{-\Delta_{\mathsf{C}}})}{\Delta_{\mathsf{C}}} \approx \frac{\Delta_{\mathsf{I}}-\Delta_{\mathsf{PAM}}}{\Delta_{\mathsf{C}}} + 1, \end{aligned} \tag{S2.11}$$

From this we see that the maximum cleavage probability is dictated by the total free-energy bias toward cleavage. The first term after the first approximate equality in the equation for n_{seed} has a simple interpretation as the point where the barrier to unbinding matches the barrier toward cleavage. For the physiological cases examined (see **Figure ??** and **??**), the values of p_{max} are between 0.7 and 1, and Δ_{C} values are order 1 as well. In this limit the second term adds a correction term that is only a small fraction of a full nucleotide position and can therefore be neglected, as done in the last step in the above equation. Equation **??** can also be arrived at through taking the large-bias limit mentioned above.

Generalized targeting rule

As we do not have the experimental data to fit multiple mismatches, we do not here perform the exact calculation of the cleavage probability for multiple mismatches. Instead we start from the fact that the physiological limit of a single mismatch was well described by the large-bias limit, and so consider also multiple mismatches in the large bias limit. If the first mismatch is outside the seed, then the second mismatch (sitting say at n_{MM2}) will dominate and balance cleavage and dissociation when

$$\frac{1}{2} \approx P_{\mathsf{clv}}(n_{\mathrm{MM2}}) \approx \frac{1}{1+\exp(-\Delta T_{n_{\mathrm{MM2}}})} \quad \Rightarrow \quad \Delta T_{n_{\mathrm{MM2}}} \approx 0$$

From equation **??** we have (assuming that $n_{\mathrm{MM2}} < N$),

$$0 \approx \Delta T_{n_{\mathrm{MM2}}} = \Delta_{\mathsf{PAM}} - 2\Delta_{\mathsf{I}} + (n_{\mathrm{MM2}}-2)\Delta_{\mathsf{C}} \quad \Rightarrow \quad n_{\mathrm{MM2}} \approx n_{\mathrm{seed}} + n_{\mathrm{pair}}, \quad n_{\mathrm{pair}} \equiv \frac{\Delta_{\mathsf{I}}}{\Delta_{\mathsf{C}}} + 1,$$

which shows that the second mismatch balances cleavage and unbinding when situated a further distance n_{pair} out from n_{seed}. For each additional mismatch added, it is easy to show that the balance point shifts a further n_{pair} bases out.

Effect of PAM recognition on target selectivity

Using equation **??** we can asses how much protection a particular non-canonical PAM site offers against cleaving the host's own genome. Letting the canonical PAM have Δ_{PAM} and the non-canonical PAM have Δ'_{PAM}, we can write the relative cleavage probability

$$p_{\mathsf{clv}}^{\mathrm{PAM}} = \frac{1+e^{-\Delta_{\mathsf{PAM}}}\left[1+\sum_{n=1}^{N} e^{-\Delta R_n}\right]}{1+e^{-\Delta'_{\mathsf{PAM}}}\left[1+\sum_{n=1}^{N} e^{-\Delta R_n}\right]} = \frac{1+e^{-(\Delta_{\mathsf{PAM}}-\Delta_{\mathsf{PAM}}^{\mathrm{crit}})}}{1+e^{-(\Delta'_{\mathsf{PAM}}-\Delta_{\mathsf{PAM}}^{\mathrm{crit}})}} \tag{S2.12}$$

where we renamed $\Delta R_n = T_n - T_1$ and introduced the critical PAM bias $\Delta_{\mathrm{PAM}}^{\mathrm{crit}}$

$$\Delta_{\mathrm{PAM}}^{\mathrm{crit}} = \ln\left[1 + \sum_{n=1}^{N} e^{-\Delta R_n}\right]. \tag{S2.13}$$

For the case of well separated PAM biases ($\Delta_{\mathrm{PAM}} > \Delta'_{\mathrm{PAM}}$), the cleavage probability has three asymptotic regimes

$$p_{\mathrm{clv}}^{\mathrm{PAM}} \sim \begin{cases} 1, & \Delta_{\mathrm{PAM}}^{\mathrm{crit}} \ll \Delta'_{\mathrm{PAM}} \ll \Delta_{\mathrm{PAM}}, \quad \text{Region c) in Figure \textbf{??}A-B} \\ \exp\left[-(\Delta_{\mathrm{PAM}}^{\mathrm{crit}} - \Delta'_{\mathrm{PAM}})\right], & \Delta'_{\mathrm{PAM}} \ll \Delta_{\mathrm{PAM}}^{\mathrm{crit}} \ll \Delta_{\mathrm{PAM}}, \quad \text{Region a) in Figure \textbf{??}A-B} \\ \exp\left[-(\Delta_{\mathrm{PAM}} - \Delta'_{\mathrm{PAM}})\right], & \Delta'_{\mathrm{PAM}} \ll \Delta_{\mathrm{PAM}} \ll \Delta_{\mathrm{PAM}}^{\mathrm{crit}}, \quad \text{Region b) in Figure \textbf{??}A-B.} \end{cases}$$

A minimal energetic model for target recognition

Now consider a extension of our minimal model where the transition between metastable state has energy biases (δ_{PAM}, δ_{C}, δ_{I}) in direct analogy with the kinetic biases (see equation **??**)

$$\Delta F_n = F_{-1} - F_n = \delta_{\mathrm{PAM}} + n_C(n)\delta_{\mathrm{C}} - (n - n_C(n))\delta_{\mathrm{I}}. \tag{S2.14}$$

Hence, all energies are measured with respect to the solution's free-energy.

2.7.3. Dissociation constant for catalytically inactive systems

Experiments on inactivated RGNs usually probe the fraction of sites bound at some late experimental time. Assuming the system has had enough time to equilibrate one typically calculates the dissociation constant, the concentration at which the bound fraction reaches half of its maximum value (second equality in **Equation ??**). This is done in analogy to a more simple two-state model that only has a bound state and an unbound state. To make this analogy within our model, we consider all molecules that are not in solution to be bound.

$$P_{\mathrm{ub}} = P_{-1}, \quad P_b = \sum_{n=0}^{N} P_n \tag{S2.15}$$

The binding rate from solution onto any sequence should be proportional to the concentration of RGN molecules. We set our δ_{PAM} within the context of the minimal model **Equation ??** at some reference concentration, at which we also calculate all free energies (ΔF's). Furthermore, in equilibrium Boltzmann statistics is valid:

$$\Delta\tilde{F}_n([\mathrm{RGN}]) = \Delta F_n - \log([\mathrm{RGN}]), \quad P_n \propto e^{-\Delta\tilde{F}_n} \tag{S2.16}$$

Taken together, the equilibrium fraction and dissociation constant are given by

$$\begin{aligned} P_{\mathrm{b}}([\mathrm{RGN}]) &= \frac{[\mathrm{RGN}]\sum\limits_{n=0}^{N}\exp\left[-\Delta F_n\right]}{1 + [\mathrm{RGN}]\sum\limits_{n=0}^{N}\exp\left[-\Delta F_n\right]} = \frac{[\mathrm{RGN}]}{[\mathrm{RGN}] + K_{\mathrm{D}}} \\ K_{\mathrm{D}} &= \frac{1}{\sum\limits_{n=0}^{N}\exp\left[-\Delta F_n\right]}. \end{aligned} \tag{S2.17}$$

For our minimal model of equation **??**, with a single mismatch is placed at position n_{MM}, equation **??** results in

$$\begin{aligned} K_{\mathrm{D}}(n_{\mathrm{MM}}) &= \frac{K_{\mathsf{max}}}{1+e^{(n_{\mathrm{MM}}-n_{\mathrm{seed}}^{\mathrm{eq}})\delta_{\mathsf{C}}}}, K_{\mathsf{max}} = \frac{e^{-\delta_{\mathsf{PAM}}}(e^{\delta_{\mathsf{C}}}-1)}{e^{N\delta_{\mathsf{C}}-\delta_{\mathsf{I}}}-1}, \\ n_{\mathrm{seed}}^{\mathrm{eq}} &= \frac{N\delta_{\mathsf{C}}-\delta_{\mathsf{I}}}{\delta_{\mathsf{C}}} + \frac{1}{\delta_{\mathsf{C}}}\ln\frac{1-e^{-(N\delta_{\mathsf{C}}-\delta_{\mathsf{I}})}}{1-e^{-(\delta_{\mathsf{C}}+\delta_{\mathsf{I}})}} \end{aligned} \tag{S2.18}$$

Note that this seed length $n_{\mathrm{seed}}^{\mathrm{eq}}$ does not in general equal its kinetic counterpart n_{seed} in equation **??**.

2.7.4. Details of fitting procedure

Since comparing relative cleavage (or binding) on constructs containing 1 mismatch (or a set of consequetive mismatches) leads to a probability/dissociation constant as in equations **??** and **??**, we fit a sigmoidal function to the data. Where replicates were available, we created 1000 bootstrapped replicates, and for each performed a straight least square fit by minimizing

$$\chi^2 = \sum_{i=1}^{N} \left(P_{\mathrm{data}}(i) - P_{\mathrm{model}}(i)\right)^2 \tag{S2.19}$$

In Figure **??** and **??**A-C, we used the bootstrapped median values for all three parameters, and report the 68% confidence intervals.

In case of the dataset from [**?**] no such replicates were available. In stead, we used the reported averages and standard deviations to minimize

$$\chi^2 = \sum_{i=1}^{N-1} \left(\frac{P_{\mathrm{data}}(i) - P_{\mathrm{model}}(i)}{\sigma_{\mathrm{tot}}(i)}\right)^2 \tag{S2.20}$$

where we had to take the finite precision of measurements in to account as some errors were reported as zero. This was done through taking

$$\sigma_{\mathrm{tot}} = \sqrt{\sigma_{\mathrm{STD}}^2 + \sigma_{\mathrm{round}}^2} \tag{S2.21}$$

with σ_{STD} being the reported statistical error amongst multiple replicates and $\sigma_{\mathrm{round}} = 0.5$ a lower estimate of the error introduced by having a finite precision in the measurement. Since the most rapid transition out of the seed region that can be recorded is over one base pair, $w_{\mathrm{min}} = 1$, we know the highest measurable Δ_{C} is $\Delta_{\mathsf{C}}^{\mathrm{max}}$ (see equation **??**). Therefore, we cannot discriminate amongst Δ_{C} values beyond 4, and we have constrained our fits to respect this condition.

2.7.5. Cascade binds its guide in sections

After assembly of the Cascade complex onto the guide RNA, every 6^{th} base is flipped out and does not interact with the target. Incorporating this into the parameterization of our model we assume that the kinetic bias does not dependent on the sequence of guide and target at these positions,

$$\delta_{\mathrm{C}}(n_{\mathrm{flip}}) = -\delta_{\mathrm{I}}(n_{\mathrm{flip}}), \quad \forall n_{\mathrm{flip}} \in (6, 12, 18, 24, 30) \tag{S2.22}$$

To perform the fit shown in Figure **??**D, we chose one particular realization of this condition with $\delta_{\mathrm{C}}(n_{\mathrm{flip}}) = -\delta_{\mathrm{I}}(n_{\mathrm{flip}}) = 0$. To allow us to fit a continuous curve to the data, data points at any of the n_{flip} positions where not taken into account and the remaining data points where re-indexed accordingly. The resulting plot shows the piecewise continuous curve when we re-introduce the flipped out bases by equating the dissociation constant to its wildtype value at these positions.

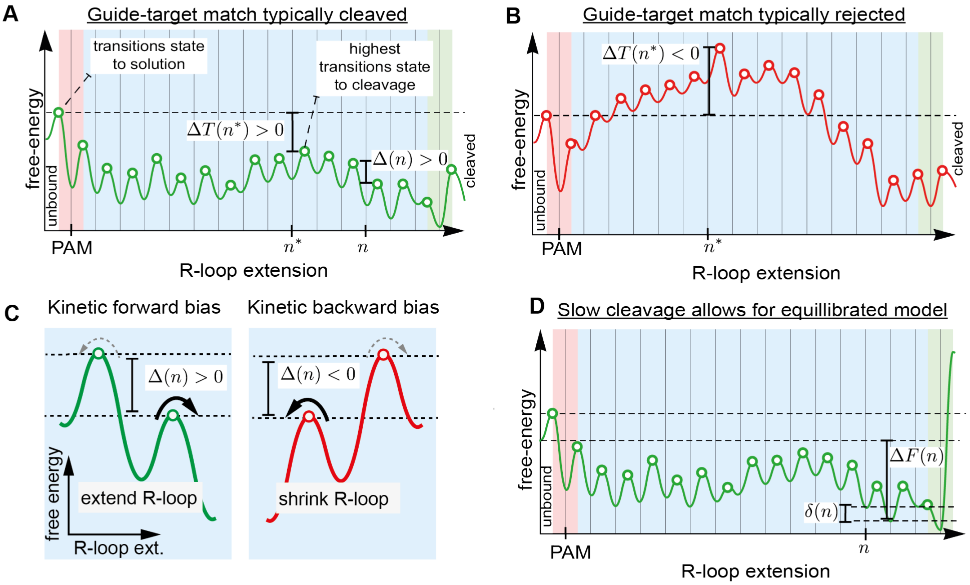


Figure S2.1: General Energy landscapes, related to figure ??. **(A)** Free-energy landscape underlying the scheme of figure 1A. Our model is completely determined by the set of transition states (open circles). The largest barrier opposing cleavage, is given by the point with the highest drawn transition state (smallest ΔT). In the limmit of large kinetic biases (see **Methods: 'high bias limit'**), it is this barrier that dominates the probality to cleave the target sequence represented. The landscape shown represents a target that is likely cleaved as the largest barrier is opposing unbinding rather then cleavage, or, in other words, the highest transition state lies below the unbinding transition (left most circle). **(B)** On the contrary, a target will likely get rejected if the highest transition state (placed at n^*) lies above the transition state towards solution. In this scenario the largest barrier obstructing cleavage is larger then the barrier hindering unbinding. **(C)** Examples of transitions that bias the RGN to extend the R-loop if the transition state to the right lies below the one to the left (left panel), or to shrink the R-loop if the transition state to the right lies above the one to the left (right panel). The difference in heights of the transition states is refered to as a 'kinetic bias'. **(D)** Free-Energy landscape as in figure A, in which parameters in equilibrium limit are indicated. Energetic biases ($\delta(n)$) are now set by the stable states within the diagram and their cumulative gain ($\Delta F(n)$) is used to calculate the dissociation constant.

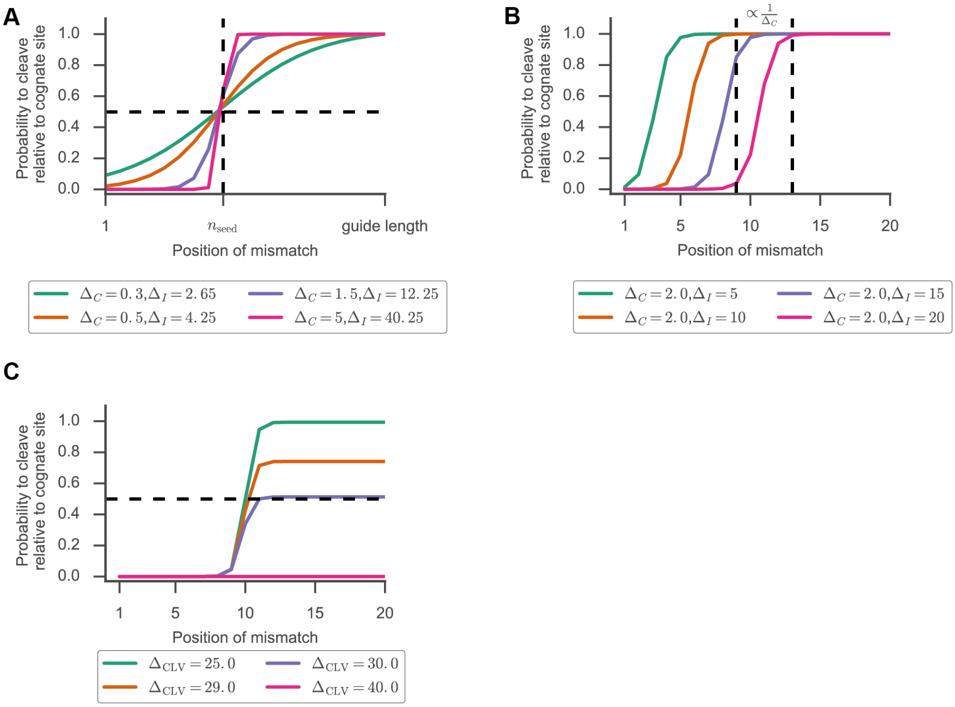


Figure S2.2: Single mismatch off-targets, related to figure ??. **(A)** Relative probability of cleaving a singly mismatched target. Seed length (n_{seed}) is kept constant by tuning Δ_C and Δ_I, while ensuring equation 2 of the main text is satisfied ($\Delta_{\text{clv}} = -100 k_B T, \Delta_{\text{PAM}} = 0.25 k_B T$). **(B)** The width of the transition region from seed to non-seed is set by the positive bias for correct base pairs (Δ_C)($\Delta_{\text{clv}} = -100 k_B T, \Delta_{\text{PAM}} = 0.25 k_B T$). **(C)** Tuning the intrinsic bias against cleavage (Δ_{clv}) allows for differential targeting of sequences with PAM distal mismatches by shifting $p_{\max}$ of equation 1 of the main text ($\Delta_\text{C} = 3 k_B T$, $\Delta_{\text{PAM}} = 3 k_B T$, $\Delta_\text{I} = 30 k_B T$).

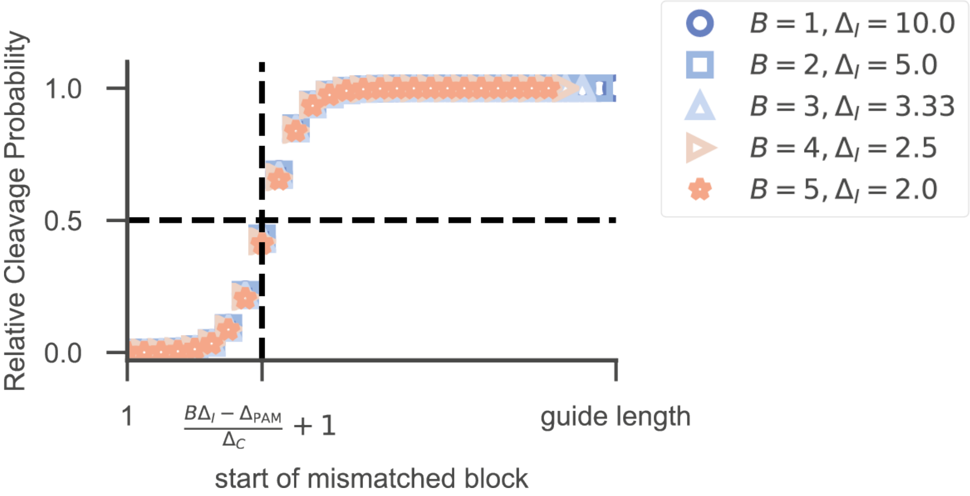


Figure S2.3: Block of mismatches, related to figure ??. The probability to cleave a target with B consequetive mismatches is equal to the probability to cleave a target with a single mismatch (placed at the start of the block) and with a mismatch bias scaled by the length of the block ($\Delta_\text{C} = 1k_BT$, $\Delta_\text{PAM} = 2k_BT$, $\Delta_\text{clv} = -100k_BT$).

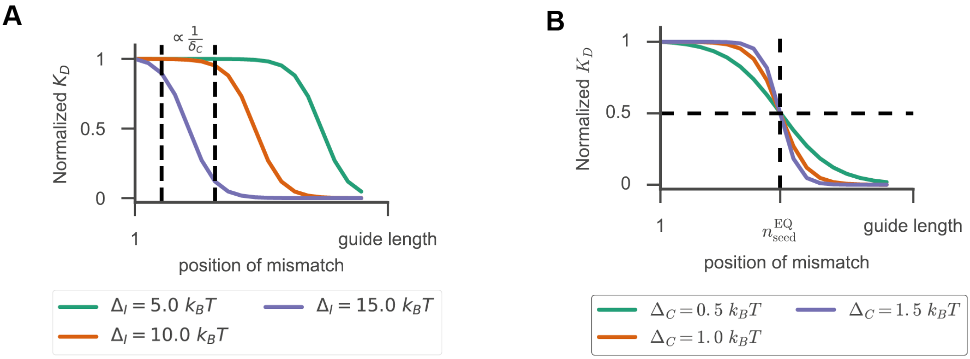


Figure S2.4: Dissociation constant for single-mismatch targets, related to figure ??. **(A)** Dissociation constant for singly mismatched targets. Fixing δ_C fixes the width of the curve, the steepness of the transition from seed to non-seed ($\delta_\text{PAM} = 3_kBT$, $\delta_\text{C} = 1k_BT$). **(B)** Fixing the ratio between match and mismatch energies fixes the seed length ($n_{\text{seed}}^{\text{EQ}}$ through equation **??**) ($\delta_\text{PAM} = 3k_BT$, $\delta_\text{I} = 10\delta_\text{C}$).

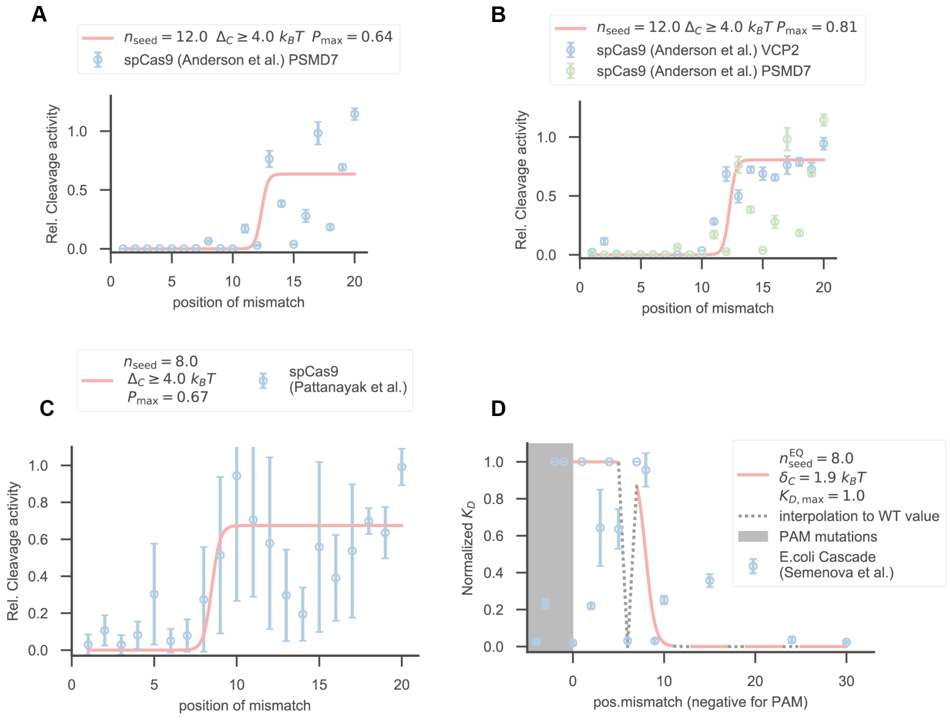


Figure S2.5: Additional comparison to experimental data , related to figure ??
(A) Data from Anderson *et al.* [**?**], PSMD7 target Confidence interval for fit parameters (68%): Δ_C[0.43,4.0), n_{seed} [12.3,14], $p_{\max}$[0.53,0.75]. **(B)** Data from Anderson*et al.* [**?**],global fit to both target sites (VCP2 target is shown in Figure 6 of main manuscript). Confidence interval for fit parameters (68%): Δ_C[0.59,4.0), n_{seed} [10.9,13.9], $p_{\max}$[0.50,1.0]. **(C)** Data from Pattanayak *et al.* [**?**], for each mutation position the median score of all single-mismatched targets within the library with the mutation at that location was used. Errorbars indicate standard deviation. Confidence interval for fit parameters (68%): Δ_C[0.20,4.0), n_{seed} [7.5,14.3], $p_{\max}$[0.58,0.98]. **(D)** Data from Semenova *et al.* [**?**], fit performed after accounting for the assembly of Cascade onto its guide in sections. All experimental data shown corresponds to mean ± standard deviation.

References

[] D. B. T. Cox, R. J. Platt, and F. Zhang, *Therapeutic genome editing: Prospects and challenges,* Nature Medicine **21**, 121 (2015).

[] J. Tycko, V. E. Myer, and P. D. Hsu, *Methods for Optimizing CRISPR-Cas9 Genome Editing Specificity,* Molecular Cell **63**, 355 (2016), arXiv:15334406 .

[] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, *Multiplex genome engineering using CRISPR/Cas systems,* Science **339**, 819 (2013), arXiv:20 .

[] G. Gasiunas, R. Barrangou, P. Horvath, and V. Siksnys, *Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria,* Proceedings of the National Academy of Sciences **109**, E2579 (2012), arXiv:arXiv:1408.1149 .

[] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, *A Programmable Dual-RNA – Guided,* Science **337**, 816 (2012), arXiv:38 .

[] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church, *RNA-guided human genome engineering via Cas9,* Science **339**, 823 (2013), arXiv:arXiv:1011.1669v3 .

[] B. Zetsche, J. S. Gootenberg, O. O. Abudayyeh, I. M. Slaymaker, K. S. Makarova, P. Essletzbichler, S. E. Volz, J. Joung, J. Van Der Oost, A. Regev, E. V. Koonin, and F. Zhang, *Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System,* Cell **163**, 759 (2015), arXiv:NIHMS150003 .

[] B. Wiedenheft, S. H. Sternberg, and J. A. Doudna, *RNA-guided genetic silencing systems in bacteria and archaea,* Nature **482**, 331 (2012), arXiv:37 .

[] C. Anders, O. Niewoehner, A. Duerst, and M. Jinek, *Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease,* Nature **513**, 569 (2014), arXiv:NIHMS150003 .

[] E. M. Anderson, A. Haupt, J. A. Schiel, E. Chou, H. B. Machado, Ž. Strezoska, S. Lenger, S. McClelland, A. Birmingham, A. Vermeulen, and A. V. B. Smith, *Systematic analysis of CRISPR-Cas9 mismatch tolerance reveals low levels of off-target activity,* Journal of Biotechnology **211**, 56 (2015).

[] B. X. Fu, R. P. St Onge, A. Z. Fire, and J. D. Smith, *Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo,* Nucleic Acids Research **44**, 5365 (2016).

[] B. X. H. Fu, L. L. Hansen, K. L. Artiles, M. L. Nonet, and A. Z. Fire, *Landscape of target: Guide homology effects on Cas9-mediated cleavage,* Nucleic Acids Research **42**, 13778 (2014).

[] Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung, and J. D. Sander, *High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells,* Nature Biotechnology **31**, 822 (2013), arXiv:NIHMS150003 .

[] P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, T. J. Cradick, L. A. Marraffini, G. Bao, and F. Zhang, *DNA targeting specificity of RNA-guided Cas9 nucleases,* Nature Biotechnology **31**, 827 (2013), arXiv:NIHMS150003 .

[] D. Kim, J. Kim, J. K. Hur, K. W. Been, S. H. Yoon, and J. S. Kim, *Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells,* Nature Biotechnology **34**, 863 (2016).

[] B. P. Kleinstiver, S. Q. Tsai, M. S. Prew, N. T. Nguyen, M. M. Welch, J. M. Lopez, Z. R. McCaw, M. J. Aryee, and J. K. Joung, *Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells,* Nature Biotechnology **34**, 869 (2016), arXiv:15334406 .

[] C. Kuscu, S. Arslan, R. Singh, J. Thorpe, and M. Adli, *Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease,* Nature Biotechnology **32**, 677 (2014), arXiv:arXiv:1208.5721 .

[] H. O'Geen, I. M. Henry, M. S. Bhakta, J. F. Meckler, and D. J. Segal, *A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture,* Nucleic Acids Research **43**, 3389 (2015).

[] V. Pattanayak, S. Lin, J. P. Guilinger, E. Ma, J. A. Doudna, and D. R. Liu, *High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity,* Nature Biotechnology **31**, 839 (2013).

[] X. Wu, D. A. Scott, A. J. Kriz, A. C. Chiu, P. D. Hsu, D. B. Dadon, A. W. Cheng, A. E. Trevino, S. Konermann, S. Chen, R. Jaenisch, F. Zhang, and P. A. Sharp, *Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells,* Nature Biotechnology **32**, 670 (2014), arXiv:NIHMS150003 .

[] J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten, and D. E. Root, *Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9,* Nature Biotechnology **34**, 184 (2016), arXiv:15334406 .

[] P. Cameron, C. K. Fuller, P. D. Donohoue, B. N. Jones, M. S. Thompson, M. M. Carter, S. Gradia, B. Vidal, E. Garner, E. M. Slorach, E. Lau, L. M. Banh, A. M. Lied, L. S. Edwards, A. H. Settle, D. Capurso, V. Llaca, S. Deschamps, M. Cigan, J. K. Young, and A. P. May, *Mapping the genomic landscape of CRISPR-Cas9 cleavage,* Nature Methods **14**, 600 (2017).

[] R. L. Frock, J. Hu, R. M. Meyers, Y. J. Ho, E. Kii, and F. W. Alt, *Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases,* Nature Biotechnology **33**, 179 (2015), arXiv:15334406 .

[] D. Kim, S. Bae, J. Park, E. Kim, S. Kim, H. R. Yu, J. Hwang, J. I. Kim, and J. S. Kim, *Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells,* Nature Methods **12**, 237 (2015).

[] F. A. Ran, L. Cong, W. X. Yan, D. A. Scott, J. S. Gootenberg, A. J. Kriz, B. Zetsche, O. Shalem, X. Wu, K. S. Makarova, E. V. Koonin, P. A. Sharp, and F. Zhang, *In vivo genome editing using Staphylococcus aureus Cas9,* Nature **520**, 186 (2015), arXiv:15334406 .

[] S. Q. Tsai, N. T. Nguyen, J. Malagon-Lopez, V. V. Topkar, M. J. Aryee, and J. K. Joung, *CIRCLE-seq: A highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets,* Nature Methods **14**, 607 (2017), arXiv:NIHMS150003 .

[] S. Q. Tsai, Z. Zheng, N. T. Nguyen, M. Liebers, V. V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A. J. Iafrate, L. P. Le, M. J. Aryee, and J. K. Joung, *GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases,* Nature Biotechnology **33**, 187 (2015).

[] Y. Lin, T. J. Cradick, M. T. Brown, H. Deshmukh, P. Ranjan, N. Sarode, B. M. Wile, P. M. Vertino, F. J. Stewart, and G. Bao, *CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences,* Nucleic Acids Research **42**, 7473 (2014).

[] E. Semenova, M. M. Jore, K. A. Datsenko, A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. J. Brouns, and K. Severinov, *Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence,* Proceedings of the National Academy of Sciences **108**, 10098 (2011).

[] F. Jiang, D. W. Taylor, J. S. Chen, J. E. Kornfeld, K. Zhou, A. J. Thompson, E. Nogales, and J. A. Doudna, *Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage,* Science **351**, 867 (2016), arXiv:arXiv:1011.1669v3 .

[] F. Jiang, K. Zhou, S. Gressel, and J. A. Doudna, *A cas9 guide RNA complex preorganized for target DNA recognition,* Science **348**, 1477 (2015).

[] M. Jinek, F. Jiang, D. W. Taylor, S. H. Sternberg, E. Kaya, E. Ma, C. Anders, M. Hauer, K. Zhou, S. Lin, M. Kaplan, A. T. Iavarone, E. Charpentier, E. Nogales, and J. A. Doudna, *Structures of Cas9 endonucleases reveal RNA-mediated conformational activation,* Science **343** (2014), 10.1126/science.1247997.

[] H. Nishimasu, F. A. Ran, P. D. Hsu, S. Konermann, S. I. Shehata, N. Dohmae, R. Ishitani, F. Zhang, and O. Nureki, *Crystal structure of Cas9 in complex with guide RNA and target DNA,* Cell **156**, 935 (2014), arXiv:NIHMS150003 .

[] Y. Xiao, M. Luo, R. P. Hayes, J. Kim, S. Ng, F. Ding, M. Liao, and A. Ke, *Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System,* Cell **170**, 48 (2017).

[] H. Zhao, G. Sheng, J. Wang, M. Wang, G. Bunkoczi, W. Gong, Z. Wei, and Y. Wang, *Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli,* Nature **515**, 147 (2014).

[] M. H. Jo, S. Shin, S. R. Jung, E. Kim, J. J. Song, and S. Hohng, *Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs,* Molecular Cell **59**, 117 (2015).

[] E. A. Josephs, D. D. Kocak, C. J. Fitzgibbon, J. McMenemy, C. A. Gersbach, and P. E. Marszalek, *Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage,* Nucleic Acids Research **43**, 8924 (2015).

[] M. Rutkauskas, T. Sinkunas, I. Songailiene, M. S. Tikhomirova, V. Siksnys, and R. Seidel, *Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection,* Cell Reports **10**, 1534 (2015).

[] W. E. E. Salomon, S. M. M. Jolly, M. J. J. Moore, P. D. D. Zamore, and V. Serebrov, *Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides,* Cell **162**, 84 (2015).

[] D. Singh, S. H. Sternberg, J. Fei, J. A. Doudna, and T. Ha, *Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9,* Nature Communications **7**, 1 (2016).

[] S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna, *DNA interrogation by the CRISPR RNA-guided endonuclease Cas9,* Nature **507**, 62 (2014), arXiv:NIHMS150003 .

[] M. D. Szczelkun, M. S. Tikhomirova, T. Sinkunas, G. Gasiunas, T. Karvelis, P. Pschera, V. Siksnys, and R. Seidel, *Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes,* Proceedings of the National Academy of Sciences **111**, 9798 (2014).

[] T. Künne, D. C. Swarts, and S. J. Brouns, *Planting the seed: Target recognition of short guide RNAs,* Trends in Microbiology **22**, 74 (2014).

[] E. A. Boyle, J. O. L. Andreasson, L. M. Chircus, S. H. Sternberg, M. J. Wu, C. K. Guegler, J. A. Doudna, and W. J. Greenleaf, *High-throughput* *biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding,* Proceedings of the National Academy of Sciences **114**, 5461 (2017).

[] D. Bikard, W. Jiang, P. Samai, A. Hochschild, F. Zhang, and L. A. Marraffini, *Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system,* Nucleic Acids Research **41**, 7429 (2013), arXiv:NIHMS150003 .

[] J. E. Dahlman, O. O. Abudayyeh, J. Joung, J. S. Gootenberg, F. Zhang, and S. Konermann, *Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease,* Nature Biotechnology **33**, 1159 (2015).

[] J. Duan, G. Lu, Z. Xie, M. Lou, J. Luo, L. Guo, and Y. Zhang, *Genome-wide identification of CRISPR/Cas9 off-targets in human genome,* Cell Research **24**, 1009 (2014).

[] B. P. Kleinstiver, M. S. Prew, S. Q. Tsai, V. V. Topkar, N. T. Nguyen, Z. Zheng, A. P. Gonzales, Z. Li, R. T. Peterson, J. R. J. Yeh, M. J. Aryee, and J. K. Joung, *Engineered CRISPR-Cas9 nucleases with altered PAM specificities,* Nature **523**, 481 (2015), arXiv:NIHMS150003 .

[] B. P. Kleinstiver, M. S. Prew, S. Q. Tsai, N. T. Nguyen, V. V. Topkar, Z. Zheng, and J. K. Joung, *Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition,* Nature Biotechnology **33**, 1293 (2015), arXiv:15334406 .

[] B. P. Kleinstiver, V. Pattanayak, M. S. Prew, S. Q. Tsai, N. T. Nguyen, Z. Zheng, and J. K. Joung, *High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects,* Nature **529**, 490 (2016), 9605103 [cs] .

[] I. M. Slaymaker, L. Gao, B. Zetsche, D. A. Scott, W. X. Yan, and F. Zhang, *Rationally engineered Cas9 nucleases with improved specificity,* Science **351**, 84 (2016), arXiv:NIHMS150003 .

[] Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio, and J. K. Joung, *Improving CRISPR-Cas nuclease specificity using truncated guide RNAs,* Nature Biotechnology **32**, 279 (2014), arXiv:29 .

[] F. A. Ran, P. D. Hsu, C. Y. Lin, J. S. Gootenberg, S. Konermann, A. E. Trevino, D. A. Scott, A. Inoue, S. Matoba, Y. Zhang, and F. Zhang, *Double nicking by RNA-guided CRISPR cas9 for enhanced genome editing specificity,* Cell **154**, 1380 (2013), arXiv:NIHMS150003 .

[] M. Haeussler, K. Schönig, H. Eckert, A. Eschstruth, J. Mianné, J. B. Renaud, S. Schneider-Maunoury, A. Shkumatava, L. Teboul, J. Kent, J. S. Joly, and J. P. Concordet, *Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR,* Genome Biology **17**, 1 (2016).

[] S. Bae, J. Park, and J. S. Kim, *Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases,* Bioinformatics **30**, 1473 (2014).

[] F. Heigwer, G. Kerr, and M. Boutros, *E-CRISP: Fast CRISPR target site identification,* Nature Methods **11**, 122 (2014).

[] K. Labun, T. G. Montague, J. A. Gagnon, S. B. Thyme, and E. Valen, *CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering,* Nucleic acids research **44**, W272 (2016).

[] I. Farasat and H. M. Salis, *A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation,* PLoS Computational Biology **12**, 1 (2016).

[] M. Khorshid, J. Hausser, M. Zavolan, and E. Van Nimwegen, *A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets,* Nature Methods **10**, 253 (2013).

[] N. Bisaria, I. Jarmoskaite, and D. Herschlag, *Lessons from Enzyme Kinetics Reveal Specificity Principles for RNA-Guided Nucleases in RNA Interference and CRISPR-Based Genome Editing,* Cell Systems **4**, 21 (2017).

[] S. H. Sternberg, B. Lafrance, M. Kaplan, and J. A. Doudna, *Conformational control of DNA target cleavage by CRISPR-Cas9,* Nature **527**, 110 (2015).

[] R. T. Leenay, K. R. Maksimchuk, R. A. Slotkowski, R. N. Agrawal, A. A. Gomaa, A. E. Briner, R. Barrangou, and C. L. Beisel, *Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems,* Molecular Cell **62**, 137 (2016).

[] S. J. Brouns, M. M. Jore, M. Lundgren, E. R. Westra, R. J. Slijkhuis, A. P. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin, and J. Van Der Oost, *Small Crispr Rnas Guide Antiviral Defense in Prokaryotes,* Science **321**, 960 (2008), arXiv:20 .

[] M. Klein, S. Chandradoss, M. Depken, and C. Joo, *Why Argonaute is needed to make microRNA target search fast and reliable,* Seminars in Cell and Developmental Biology **65**, 20 (2017).

[] C. Xue, N. R. Whitis, and D. G. Sashital, *Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity,* Molecular Cell **64**, 826 (2016), arXiv:15334406 .

[] M. A. Nowak, *Evolutionary Dynamics: exploring the equations of life* (Harvard University Press, 2006).

[] D. P. Bartel, *MicroRNAs: Target Recognition and Regulatory Functions,* Cell **136**, 215 (2009), arXiv:0208024 [gr-qc] .

3

Mechanistic modeling explains dCas9 binding and Cas9 cleavage dynamics

Genome engineering using the RNA guided DNA endonuclease CRISPR-Cas9 is on the rise. When loaded with a single-guide RNA (sgRNA), the Cas9-sgRNA binds and cleaves the DNA site complementary to the supplied guide sequence. Unfortunately, Cas9-sgRNA is known to also cleave DNA sites with non-perfect complementarity, a phenomenon more commonly known as off-targeting. Towards quantifying the risks of its implementation, we model the (off-)target binding, dsDNA unwinding, and cleavage by Cas9-sgRNA to tell the fraction of cleaved DNA when subjected to a fixed nuclease concentration for a given time. Within the same physical model, we also capture the binding dynamics of catalytically 'dead' dCas9 and rationalize the large disparity in off-targeting observed with its active counterpart. Using a series of recent high-throughput biophysical experiments, we extract the microscopic free-energy landscape that underlies the interactions between Cas9-sgRNA and an (off-)target DNA. We reveal the major conformational change, which repositions Cas9's nuclease domains, initiates simultaneously with DNA unwinding, only to be completed once a (near) complete RNA-DNA hybrid is formed. Finally, by direct comparison and using the free-energy landscape, we rationalize how our kinetic model improves upon existing thermodynamic models.

manuscript in preparation

3.1. Introduction

CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CIRPSR associated) systems, CRISPR-Cas9 systems in particular, have opened the door to a multitude of gene editing applications [**?** **?**]. Cas9 uses two RNA molecules – the CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA) – as a guide to bind and cleave complementary (double-stranded) DNA. Most biotechnological applications instead load Cas9 with a synthetic singe-guide RNA (sgRNA) containing a 20 nucleotide (nt) long sequence designed to be complementarity to the DNA sequence one wishes to target [**?**]. The relative ease by which Cas9-sgRNA can be programmed to bind and cleave any (genomic) DNA sequence of interest has enabled its use in gene silencing/activation [**?**], fluorescent imaging of genomic loci [**?**], RNA or DNA detection [**?** **?**] and genome editing [**?** **?**].

Structural [**?** **?**] and biophysical [**?** **?** **?**] studies indicate that Cas9's two nuclease domains (HNH and RuvC) are activated only after binding the DNA target, which is often taken to imply Cas9 is reasonably specific. However, Cas9-sgRNA also targets sites (off-targets) other than those fully complementary to its guide (the on-target) [**?** **?** **?** **?** **?** **?** **?** **?** **?**]. Such off-targeting can induce unwanted genomic alterations, including point mutations, large-scale deletions or chromosomal rearrangements [**?**]. Due to the high risk of deleterious effects, such editing errors have impeded a wide-spread implementation of Cas9-sgRNA in human therapeutics.

Though experiments have demonstrated that the position of mismatches along the guide-to-target hybrid strongly influences both binding and cleavage activities, the process behind this is not yet quantitatively understood. For example, catalytically inactive ('dead') dCas9 notoriously binds more off-targets sites than Cas9 cleaves [**?** **?** **?** **?**], and there is at present no way of translating binding affinities into cleavage propensities, or vise versa.

Here we unify binding and cleavage of *Streptococcus pyogenes* Cas9 (spCas9) within a single kinetic model. We expect such a physics-based framework to hold several advantages compared to existing *in silico* prediction tools that are either based on empirically derived scoring schemes [**?** **?**] or Machine Learning approaches [**?** **?**] utilizing scoring schemes derived and hidden within a "black box" algorithm. First, all our model parameters are physically interpretable, rates and energies determining the binding/cleavage reactions. As a result, the model's output is physically interpretable as well, returning effective reaction rates for either binding or cleavage reactions under variable experimental conditions. This allows us to tell more than what off-targets are cleaved most (at steady-state) and answer the question: *"What fraction of my off-target pool is bound or cut at a given nuclease concentration and after a given time?"* Hence, such a model offers an *in silico* testing-ground for future binding or cutting based experiments.

Second, using the language of free-energy landscapes allows us to tie reaction intermediates (metastable states) to structural data.

Expanding upon our own kinetic modeling efforts (**Chapter ??**)[**?**] we shall use three high-throughput biophysical datasets to elucidate the free-energy landscape that (d)Cas9-sgRNA experiences while interacting with (off-)target DNA. First, Boyle et al. [**?**] measured the rate of change in bound DNA fraction at fixed dCas9-sgRNA concentration in the first 1500 seconds of the reaction for a library of off-targets. Second, Jones et al. [NucleaSeq data from [?]] used saturating concentrations of (active) Cas9-sgRNA to determine effective cleavage rates. Finally, Jones et al. [CHAMP data from [?]] independently measured the half-saturating concentrations 10 minutes after introducing (inactive) dCas9-sgRNA. We demonstrate that our parameterized model is capable of accurately describing all three quantities. Moreover, we can predict the half-saturating concentrations, while training the model only with data taken at fixed concentration. To the best of our knowledge, we thereby present the first physical model capable of quantitatively describing both binding and cleavage reactions, for both varying (d)Cas9-sgRNA concentrations and incubation times.

The free-energy landscape we propose, the extracted model parameters, helps us explain experimental observations in terms of reaction rates for the sub-processes of initial target binding, (partial) hybrid-formation and inducing the DNA breaks. In particular, the free-energy landscape helps us understand how Cas9 balances being both an efficient (high enough activity on on-target) and specific (low enough activity on off-targets) nuclease, at the cost of binding more promiscuously. We show mismatches come at (nearly) equal energetic costs throughout the guide-target hybrid, while the free-energy representing interactions with the on-target shows a distinct position dependence. We shall demonstrate how the previously characterized conformational rearrangements involving Cas9's two nuclease domains [? ? ?] manifests itself within our proposed Cas9-gRNA free-energy landscape. Hence, we thereby unify observations across bulk and single-molecule experiments.

Finally, we demonstrate how both the state-of-the-art prediction tool [?], as well as the recently published model by Zhang et al. [?], can both been seen as a limiting case of our more general model. By direct comparison of predictions and by showing that we are not in the required limits, we shall explain exactly how our model improves upon the existing ones.

3.2. Results

3.2.1. A kinetic model for target recognition by (d)Cas9-sgRNA

The reaction scheme underlying our model is shown in **Figure ??**A. A Cas9-sgRNA from solution binds a DNA target after first using protein-DNA interactions to recognize a 3nt 'protospacer adjacent motif' (PAM) sequence – canonically 5'-NGG-3' – located on the non-target DNA strand [? ?]. Binding to the PAM triggers a conformational change that enables interactions with the +1 DNA base pair [? ?] initiating sequential formation of a DNA-Cas9-sgRNA-DNA 'sandwich', called the R-loop [? ? ? ?]. The R-loop can grow and shrink until unbinding or reaching completion, after which Cas9 uses its two nuclease domains (HNH and RuvC) to cleave the target and non-target DNA strands [?].

While existing theoretical models only incorporate the thermodynamics [? ?], we (**Chapter ??**)[?] and others [?] have emphasized the importance of incorporating the kinetics of the PAM binding, hybridization and cleavage reactions to explain several experimental observations. To build a kinetic model of target recognition by Cas9-sgRNA, we treat every

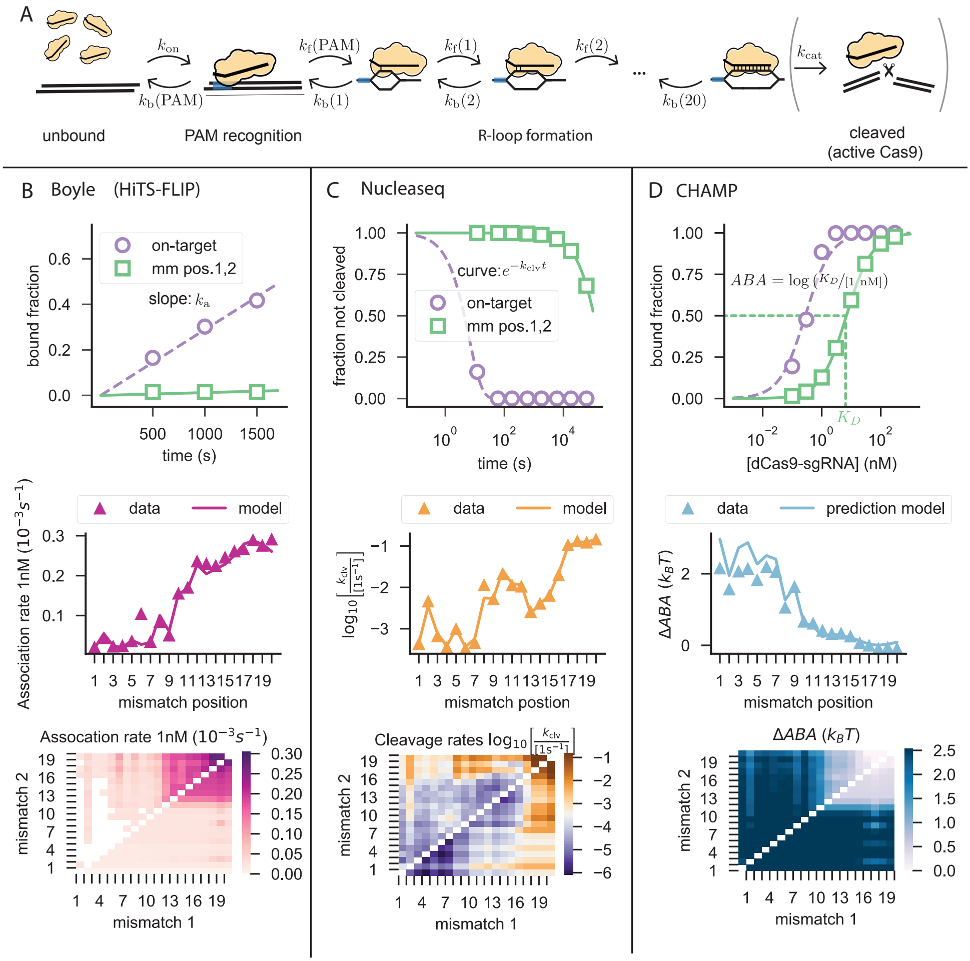


Figure 3.1: kinetic model captures both binding and cleavage data. **(A)** General reaction schema underlying our kinetic model. A Cas9-sgRNA from the soluble pool (with known concentration) binds the DNA at the PAM site, sequentially progresses through R-loop formation, and eventually cuts the DNA. The set of forward and backward rates describing transitions (arrows) between states (images) fully parameterize our model. **(B)** Fit to HiTS-FLIP data [**?**]. top: the association rate (k_a) is estimated as the slope of a straight line forced through the origin and fitted to three measurement points (see **S.I.**). Figure here shows representative calculations using the extracted model parameters. middle: fit against off-targets with 1 mismatch. Bottom: fit against off-targets with 2 mismatches (data in upper triangle/ model in lower triangle). **(C)** Fit to NucleaSeq data [**?**]. top: the cleavage rate (k_{clv}) is estimated by an exponential fit to the fraction of uncut off-target DNA (see **S.I.**). middle: fit against off-targets with 1 mismatch. Bottom: fit against off-targets with 2 mismatches (data in upper triangle/ model in lower triangle). **(D)** Prediction of CHAMP data [**?**]. top: ABA values are the logarithm of the half-saturation concentration after 10 minutes of dCas9-sgRNA interactions with DNA (see S.I.). middle: prediction of off-targets with 1 mismatch. Bottom: prediction of off-targets with 2 mismatches (data in upper triangle/ model in lower triangle).

intermediately sized R-loop (1,2,...,20 nt) as well as the PAM bound and unbound (solution) configurations as metastable states, and transitions between states as being thermally activated. In general, the model is completely parameterized by the set of forward and backward rates (**Figure** 1A) for every Cas9-sgRNA-DNA combination.

Though the type of mismatch matters, experimental data also shows consistent trends in both binding and cleavage activity with respect to the position for any mismatch type (i.e. the data shown in this work). As a step towards a full sequence dependent model, and to uncover any sequence independent determinants of targeting activity, we here use a (target) sequenced averaged parameterization of a specific Cas9-sgRNA combination. In this scenario, all internal forward reactions represent the same process of removing one DNA-DNA base pair and forming a new one between the guide RNA and target strand DNA. To simplify matters, we assume only backward reactions can be dependent on position and the complementarity between RNA and DNA, thereby carrying all position dependency. This assumes the transition state encountered when extending the R-loop occurs before the RNA base interacts with the DNA base (the dsDNA always matches). Hence, apart from a (concentration dependent) rate of binding from solution onto the PAM (k_{on}) and the final rate of inducing the DNA breaks (k_{cat}), a single forward rate (k_{f}) is used to parameterize all remaining forward reactions (**Figure ??**A, **S.I.**). Although no direct evidence that forward rates must be position independent, we shall show that the current parameterization is sufficient to capture the trends in the data. Instead of using backward rates as our model parameters directly, we use the detailed balance condition ($k_{\text{b}}(n) = k_{\text{f}}e^{F_{\text{n}}-F_{\text{n-1}}}$) to relate every backward rate to the forward rate and the difference in free-energy between consecutive states ($F_{\text{n}} - F_{\text{n-1}}$, **Figure ??**A, **S.I.**). As we assume that placing a mismatch at the position within the R-loop promotes only the corresponding backward rate, this implies all free-energies from the position onwards will be raised by the same amount (**S.I.**).

All in all, a total of 44 independent parameters describe target binding and cleavage of a fixed Cas9-sgRNA at any DNA target: (1-2) The rate of PAM binding from solution, k_{on}, and the free-energy gained/lost in this process, F_{PAM} (both at the (d)Cas9-sgRNA concentration the data is taken), (3) the forward rate k_{f}, (4-23) 20 free-energy differences describing progressing the R-loop when guide and target are matching, (24-43) 20 penalties for mismatches within the R-loop that (locally) increase the difference in free-energy, raising the on-target's landscape from the position of the mismatch onwards, and (44) the catalytic rate k_{cat} which is set to zero when considering nuclease inactive dCas9 [see **Figure ??**A, **S.I.** for details].

As Cas9 is known to interact with the DNA, especially with the non-target strand [**?**], the target recognition process is not fully described by the hybridization energies of the nucleic acids alone. For this reason, adding a matching base pair to the hybrid does not need to be energetically favorable, and the parameters corresponding to matches can include any form of protein-DNA interactions or conformational changes that couple to R-loop progression. Mismatch penalties are assumed to be positive, as replacing a match with a mismatch is by definition energetically unfavorable.

3.2.2. Modeling measurable quantities for both dCas9 and Cas9

We have set it as our goal to quantitatively describe the outcome of both binding and cleavage experiments within a single physical framework. To this end, three independent high-throughput biophysical datasets were used to compare against our model.
First, Boyle et al. [**?**] used a high-throughput fluorescence microscopy assay (HiTS-FLIP: 'high-throughput sequencing-fluorescent ligand interaction profiling') to determine the rate of change in the bound DNA population (for a large library of off-targets) within the first 1500 seconds upon introducing dCas9-sgRNA (top panel **Figure ??**B). We used a master equation formulation to numerically determine the temporal evolution of the bound fraction at any off-target, which we interpreted as the equivalent of (background corrected) fluorescence intensities. From here, we extracted the reported (effective) associate rate (k_{a}) by mimicking the procedure used in the experiments by Boyle et al. (top panel **Figure ??**B, see **S.I.** for details). Note this effective association rate does not equal the binding rate from solution (k_{on}), but rather is modulated by the rate of rejection from the DNA, explaining its dependence on mismatch configuration.

A second experiment, the CHAMP ('chip-hybridized association- mapping platform') assay [**? ?**], similar to HiTS-FLIP, uses a high-throughput fluorescence setup to determine binding activities. However, while HiTS-FLIP tracks the bound fraction over time at a fixed dCas9-sgRNA concentration of 1nM, CHAMP measures the bound fraction after a fixed time of 10 minutes for a series of concentrations. Hence, while both reporting on dCas9 binding off-targets, the CHAMP and HiTS-FLIP datasets probe the binding activity's response to uniquely varying experimental conditions. Using the bound fractions, CHAMP determines the half-saturation concentrations (effective dissociation constants) after 10 minutes of dCas9 exposure. Comparing this to a reference of 1 nM, allows one to define an 'Apparent Binding Affinity' (ABA, $\Delta\mathrm{ABA} = \mathrm{ABA} - \mathrm{ABA}_{\text{on-target}}$) as the logarithm of the relative dissociation constant (**Figure ??**D top panel, see **S.I.** for details).

Finally, Jones et al. also present the NucleaSeq (nuclease digestion and deep sequencing) technique [**?**] to measure the (effective) cleavage rates for a library of off-targets (k_{clv}) by monitoring the fraction of uncut DNA over time and fitting this to a single exponential function (top panel **Figure ??**C). The **S.I.** shows how we numerically determined k_{clv} for all off-targets within the experimental library. Note that is not the same as the intrinsic catalytic rate (k_{cat}) we have set as a model parameter. Rather, $k_{\mathrm{clv}} \leq k_{\mathrm{cat}}$, as NucleaSeq reports the (inverse) average time to bind the target, complete the R-loop and induce the DNA breaks (which happens at the rate k_{cat}), explaining how k_{clv} can depend on the off-target sequence.
All three experiments used the same guide sequence derived from λ-phage DNA. (CHAMP and NucleaSeq additionally used the same off-target library), thereby minimizing potential sequence dependencies that would effect a successful translation between the datasets by our model.

As a first approach we have fit our model against the HiTS-FLIP data alone, leaving the others as tests (**Figure ??**). **Figure ??**A shows the fit against all library members with a single mismatch (top panel), and those with two mismatches (bottom panel), together forming the entire dataset used to fit. **Figure ??**B shows that we captured measured values of CHAMP with high precision (top panel: one mismatch, bottom panel: two mismatches, combined correlation coefficient: 93%). This strongly indicates that the model's implementation of varying nuclease concentration is valid, as this prediction of dCas9 binding at varying concentrations (CHAMP) is based on model parameters extracted solely from the HiTS-FLIP data taken at 1 nM. Further, using a $k_{\text{cat}} \gg 1s^{-1}$ to not make this the rate limiting step, we predict cleavage rates from NucleaSeq with approximately 84% correlation, again only using dCas9 based information (**Figure ??**C shows comparison to library members with one mismatch on top and with two mismatches below). Yet, **Figure ??**C reveals that the model underestimates k_{clv} for many off-targets (which cannot be resolved by a further increase in k_{cat}). In addition, our stochastic optimization algorithm (see *S.I.* for details) returned relatively strongly varying parameter sets, while still giving similar fit qualities (**Figure** S1D). **Figures ??**D-F show the parameter set (**Figure ??**D: on-target free-energy landscape, **Figure ??**E: mismatch penalties, **Figure ??**F: rate parameters) of the best fit (lowest χ^2, see **S.I.**) that was used to produce **Figures ??**A-C together with parameter sets belonging to fits that differ less than 5% in their prediction of the fitted HiTS-FLIP data (see **S.I.**). We noticed that apart from the on-target's free-energy at the PAM and 11-12 nt into the R-loop, most parameters are allowed to vary significantly without apparent loss in fit quality. Especially the strongly varying mismatch penalties (**Figure ??**E) and rate parameters (**Figure ??**F) may not affect the resulting association rates (**Figure ??**A), but strongly affect the cleavage rates (**Figure ??**C). In the coming section we shall describe the obtained parameters in more detail. For now, we note that fitting our model only to association rates can constraint our model parameters enough to describe CHAMP, but not enough for NucleaSeq.

We take the heterogeneity of the fit parameters (**Figures ??**D-F) as a sign that the best fit represents an overfit to the HiTS-FLIP data, capturing noise, thereby limiting our predictive power of the NucleaSeq data. In an attempt to combat this, and more confidently report the underlying kinetic parameters, we proceeded by using a simultaneous fit to HiTS-FLIP (k_{a}) together with NucleaSeq (k_{clv}) (**Figures ??**B-C, see **S.I.** for details). We reasoned that as k_{clv} values report the time needed for Cas9-sgRNA to make it from the solution state all the way through the free-energy landscape into the post-cleavage state, the prediction of the NucleaSeq data should be more sensitive to the value of the mismatch penalties and forward rates. These parameter values set the placement, height and typical crossing times of (effective) energetic barriers within the off-target free-energy landscapes. Adding this information to that coming from HiTS-FLIP, presumably being most sensitive to the stability of different states as this determines whether or not binding will be long enough lived to be observed, should be enough to constraint our model parameters sufficiently. **Figure ??** shows fit parameters, in particular the mismatch penalties up until nt 16 (**Figure ??**B) and the forward rates (**Figure ??**C) are now more strongly constrained. The combination of having typical cleavage times (NucleaSeq) at saturating conditions together with typical times to reach stable binding (HiTS-FLIP) at a fixed concentration, also strongly constrained the fitted binding rate (k_{on}) (**Figure ??**C).

More importantly, using this combined fit we see it is possible to quantitatively capture both dCas9 binding and Cas9 cleavage dynamics within a single physical framework (**Figures ??**B and C). Clearly, our model nicely reproduces values for off-targets with one (middle panel **Figure ??**B) and two mismatches (bottom panel **Figure ??**B). Using the fit to library members with up to two mismatches, we also accurately reproduce measured k_a's for all off-targets in the library with more mismatches, leading to a combined correlation of 89% (**Figure ??**A). Similarly, our model accurately reproduces cleavage rates from NucleaSeq both for single-mismatched (middle panel **Figure ??**C) and double-mismatched off-targets (bottom panel **Figure ??**C), with high accuracy (combined correlation of 93%, **Figure ??**B). Interestingly, the model recovers that a mismatches between nt 12 and nt 17 can strongly reduce cleavage activity (**Figure ??**C, middle panel) while minimally influencing apparent binding activity (**Figure ??**B, middle panel). We shall discuss the physics underlying this below. Finally, without fitting any parameters, we manage to accurately translate from the temporal sweep of HiTS-FLIP (**Figure ??**B) to the Cas9-sgRNA concentration sweep of CHAMP for all given off-targets (95% correlation, **Figures ??** and **??**C).

Taken together, we build and parameterized (as we shall discuss using **Figure ??**) a single kinetic model (**Figure ??**A) that explains the dynamics of (d)Cas9-sgRNA-DNA interactions both at various times and concentrations. Next, we shall take a further look at the physical properties of Cas9 extracted from the data and describe their consequences.

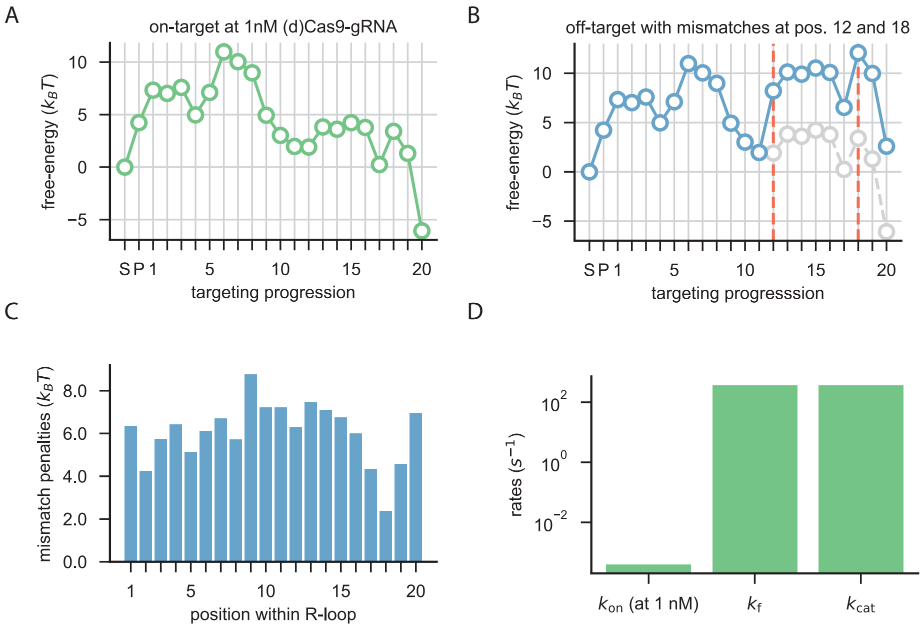


Figure 3.2: Kinetic parameters. **(A)** Free-energy landscape representing on-target DNA interacting with 1nM Cas9-sgRNA. **(B)** Free-energy landscape representing off-target DNA (mismatches at positions 12 and 18) interacting with 1nM Cas9-sgRNA (blue). On-target free-energy landscape shown in grey. **(C)** Mismatch penalties as a function of location within RNA-DNA hybrid. **(D)** Forward rate parameters.

3.2.3. Free-energy landscape of (d)Cas9-sgRNA-DNA

From a simultaneous fit (**S.I.**) to the data shown in **Figures ??**B and C, we obtain the free-energy landscape that the Cas9-sgRNA experiences upon interacting with a given DNA target, for the guide common to both experiments. **Figure ??**A shows the resulting free-energy landscape for interacting with the on-target, while **Figure ??**B shows both the on-target's landscape (grey, dashed line) together with an example landscape encountered at an off-target with mismatches placed at nt 12 and nt 18 (blue, solid line). The latter is obtained by raising all points from the 12th position onwards in the on-target's landscape by the 12th mismatch penalty, and all points from the 18th position onwards by the 18th penalty (**Figure ??**C, **S.I.**). **Figure ??**D shows the obtained rate constants.
Remarkably, the on-target free-energy landscape (**Figure ??**A) shows a distinct position dependence, which we have found to be responsible for many features seen in the dataset(s). Starting from the PAM bound state, the free-energy strongly increases and remains relatively high for the first 8 nt. Destabilizing the first 8 R-loop associated states results in an effective barrier that must be bypassed before a stable binding intermediate is reached. As a result, adding a single mismatch within this region makes the effective barrier nearly insurmountable within the time of a typical experiment. Hence, we recover what is commonly referred to as the 'seed' region wherein a single mismatch can completely disrupt either binding or cleavage [**? ?**]. The end of the seed-region contains another (slighter) increase (see nucleotides 6 to 8). Although no direct evidence, we hypothesize such an additional barrier reflects the cost of rearranging the guide outside the seed into proper helical form to enable further hybrid formation [**?**].
After the unstable seed, the bound state gradually becomes more stable when forming nt 10-12, reaching a local minimum after the 12th base pair. Interestingly, before reaching a final cleavage competent state (full R-loop), the free-energy landscape reveals a second effective barrier after nucleotide 13. Below we shall show the presence of two regions of unfavorable R-loop progression is consistent with experimentally established conformational dynamics of Cas9's nuclease domains.

The mismatch penalties (**Figure ??**C) remain rather constant (at about $6 \pm 1 k_B T$) throughout. Notable exceptions are nucleotides 2, 9 and those from 17 until 19. The lower mismatch penalty of around 4 $k_B T$ at the second R-loop position originates the increased activity seen for both dCas9 and Cas9 when muting nt 2 compared to mutating either of its neighbors (**Figures ??**B-D). Similarly, as placing the first of two mismatches at the 9th position results in a lower cleavage rate compared to placing it at either the 8th or 10th position, we fit an increased mismatch penalty of around 9 $k_B T$. Mutating nucleotides 17-19 comes at a lesser cost of 4 $k_B T$, compared to most of the other positions. This, together with the on-target target binding being always more stable than initial PAM recognition after the 17th base pair (**Figure ??**A), is consistent with a previous reports that have shown Cas9 can indeed cleave substrates that contain mismatches at nucleotides 17-20 with only slight hindrance [**? ?**].
The fitted rate constants of **Figure ??**D reveal that, at 1nM Cas9-sgRNA, PAM recognition happens at a rate (k_{on}) that is 5 orders of magnitude less than the rate of progressing the R-loop (k_{f}) and the rate of catalyzing cleavage (k_{cat}). The large forward rate (k_{f}) results in similarly high rates for shrinking the R-loop ($k_{\text{b}}(n) = k_{\text{f}} e^{F_{\text{n}} - F_{\text{n-1}}}$, see **S.I.**). Yet, despite

growing or shrinking the R-loop by one nucleotide happening rather fast, the shear amount of such steps needed before the full R-loop is formed makes that R-loop formation is still the rate limiting process to cleavage (and stable binding), thereby governing Cas9's mismatch tolerance.

In conclusion, our physical model allows us to the extract the free-energy landscape describing the interaction between target DNA and a Cas9-sgRNA complex. In what follows, we shall first in more detail explain how the landscape shown in **Figure ??**A captures Cas9's major conformational change, and show how this results in the pronounced difference between binding (dCas9) and cleavage (Cas9) activities

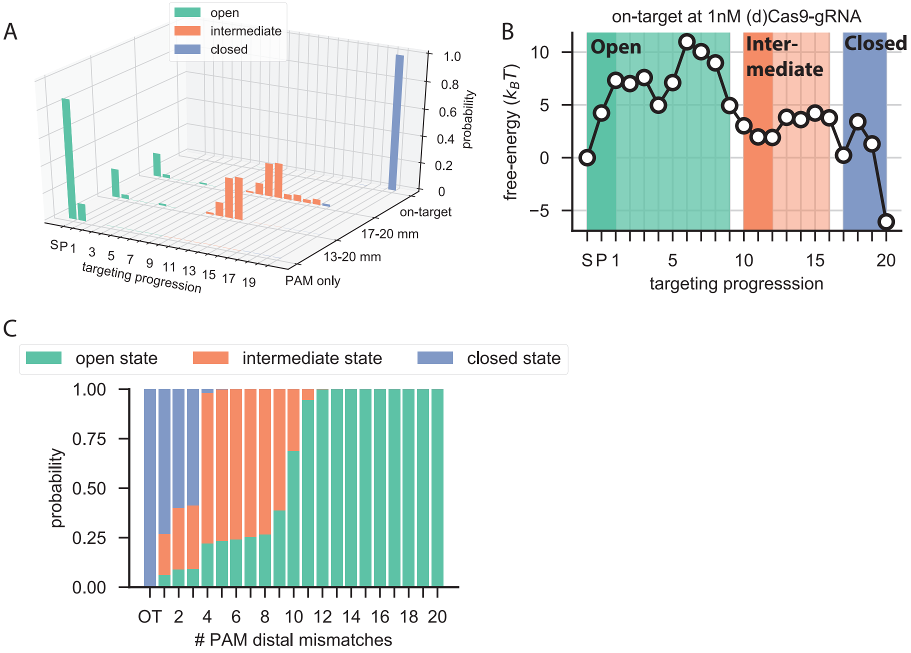


Figure 3.3: Relating free-energy landscape to Cas9's conformational dynamics. **(A)** Equilibrium occupancies (10nM dCas9-gRNA) for all 21 microscopic states, and different off-targets. This mimics the FRET histograms shown in Figure 1C of [**?**] **(B)** A coarse-grained view of the on-target free-energy landscape (**Figure ??**A). Using the nomenclature of [**?**] we identify the 'open', 'intermediate' and 'closed' states. Solid colors show the states with the greatest contribution (the most stable states in every subgroup). **(C)** A. Fraction of equilibrium occupancies for each of the three coarse-grained states, shown for off-targets with increasing number of consecutive PAM distal mismatches.

3.2.4. Conformational change of Cas9's HNH domain couples to R-loop formation

Figure ??A reveals that although forming a complete R-loop with the on-target (at 1nM) is energetically favorable, reaching this cleavage competent state is preceded by surpassing two regions of significant instability. This is surprising, given we have previously showed

(**Chapter ??**)[**?**] that the sequential nature of the R-loop formation process in itself dictates clear mismatch position dependent unbinding/cleavage rates at off-targets. Even with a constant gain for every match added to the R-loop, the placement of a mismatch still modulates the barrier opposing rejection of the off-target. As the free-energy landscape of **Figure ??**A clearly deviates from one with a constant downward slope, it must be the result of structural properties of the Cas9 protein that couple to hybrid formation.

A comparison of guide-bound and target-bound structures revealed Cas9 undergoes a conformational change in which the active sites of its HNH and RuvC nuclease domains are repositioned favorably for cleavage [**? ? ?**]. A bulk FRET experiment, in which two of dCas9's (initially distant from each other) amino-acids are fluorescently labelled, confirmed the HNH domain rearranges itself prior to cleavage, and showed the RuvC domain movement is strictly coupled to that of the HNH domain [**?**]. More recently, single-molecule FRET studies have shown the existence of two dominant bound configurations of Cas9-sgRNA [**? ? ?**]. As the HNH domain moves, the distance between the fluorescent dyes changes, resulting in an altered FRET efficiency. By collecting the FRET efficiency traces of many molecules, observed for long enough time, one obtains an estimate of the equilibrium occupation in the state space along the FRET coordinate, the position of the HNH domain.

Given the free-energy landscapes for both on-target and off-targets (**Figure ??**), we can directly calculate the equilibrium dCas9 occupation in each state according to the Boltzmann distribution (**S.I.**), which is what the FRET efficiency histograms attempt to estimate. **Figure ??**A displays equilibrium distributions for various amounts of PAM distal mismatches, thereby directly mimicking the experiment performed by Dagdas et al. (see Figure 1C in [**?**]). In line with the authors' findings, we confirm dCas9-sgRNA-DNA is mainly found in one of three states (conformations) (indicated by different colors in **Figure ??**B). When subjected to on-target DNA, nearly all bound molecules are cleavage competent (occupying the final state). Introducing mismatches causes dCas9-sgRNA-DNA to get trapped in an intermediate configuration (the orange colored peak around nt 12 in **Figure ??**A). Four or more terminal mismatches is sufficient to effectively deplete the final state (blue bars in **Figure ??**A). As the target contains more mismatches, the initial (bound) state (the peak seen for the solution and PAM states in **Figure ??**A) becomes more favorable. **Figure ??**C shows the fractions of molecules occupying each of the three 'coarse-grained states' (defined in **Figure ??**B) as a function of the number of consecutively placed PAM distal mismatches. Using the terminology introduced by Yang et al. [**?**], we identify an 'open' HNH conformation (roughly corresponding to the microscopic states up until the 8th base pair in **Figure ??**A), a 'closed' configuration (roughly corresponding states 17–20 in **Figure ??**A), as well as an 'intermediate' configuration (states 9–16 in **Figure ??**A). In agreement with the study of Dagdas et al. [**?**], the system gradually switches from mainly occupying the closed state, to the open state as more mismatches are introduced, transiting via the intermediate state in the process. We note the smFRET studies probe the reaction coordinate along the HNH conformational change, whereas our model's reaction coordinate indicates targeting progression (PAM binding + R-loop formation). The similarity between our model and the data discussed here thus reveals a likely equivalence of these two point of views. We conclude that Cas9's nuclease domains must rearrange themselves in order for the R-loop to extend. Moreover, **Figure ??**B shows that the conformational change is split over two major barriers, with the first barrier being encountered straight after binding to the PAM.

Furthermore, Yang et al. mention that although three main FRET values were observed on any (off-)target, the value of the intermediate state depends on the number of mismatches introduced – signifying the HNH domain adopted a (slightly) different configuration. Indeed, **Figure ??**A shows that with 7 PAM distal mismatches the R-loop is unlikely to progress passed the 12th base pair, while the likelihood of observing a partial R-loop of length 16 is many times higher with only 4 mismatches, both corresponding to what we identify as 'the intermediate HNH state' in **Figure ??**B. The reported shift in FRET value upon introduction of more mismatches is consistent with our model's prediction that Cas9-sgRNA-DNA occupies different microscopic states. This is in line with our finding that the conformational change happens throughout the hybrid formation process.

Finally, we note that only the closed state is found to be cleavage competent [**?**], also consistent with our model. We conclude that the free-energy landscape (**Figure ??**) obtained by fitting bulk data (**Figure ??**) is not only consistent with, but complements structural and single-molecule data on (d)Cas9.

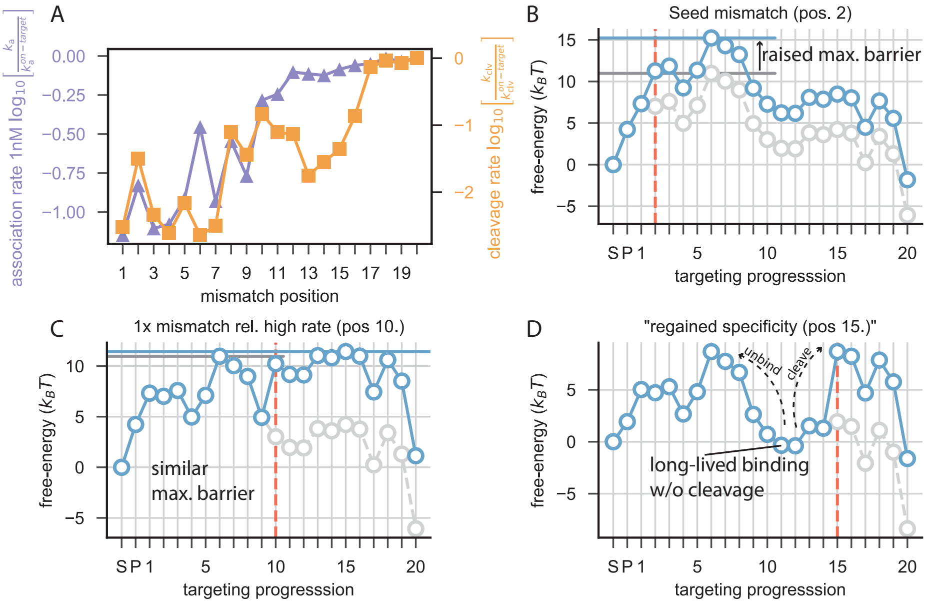


Figure 3.4: Difference between binding and cleavage activities. **(A)** Association rates from HiTS-FLIP (purple triangles) and cleavage rates from NucleaSeq (orange squares), for single-mismatch off-targets, both normalized to the corresponding on-target rates. **(B)** Free-energy landscape for off-target (mismatch at position 2) (blue) together with on-target (grey). A seed mismatch significantly raises the largest barrier (horizontal lines) opposing both binding and cleavage. **(C)** With a mismatch at position 10, binding and cleavage still are limited by the same barrier (horizontal line). Compared to placing a mismatch in the seed (figure B), the off-target landscape (blue) is raised far less in comparison to the on-target landscape (grey). **(D)** mismatch at position 15 causes binding and cleavage to be limited by different barriers. Binding is stabilized after surpassing the first barrier (entering position 12), whereas cleavage requires Cas9 to surpass also the second barrier visible.

3.2.5. Promiscuous binding helps Cas9 to be both a specific and an efficient nuclease

With the free-energy landscape in hand, we can now explain what off-target sequences typically lead to binding without cleavage. **Figure ??**A overlays the data from Nucleaseq (orange squares) and HiTS-FLIP (purple triangles) experiments for singly-mismatched off-targets, both normalized to their respective on-target values. Clearly, placing just a single mismatch within approximately 8 nucleotides from the PAM significantly slows down both binding and cleavage. **Figure ??**B shows the free-energy landscape for a target with a mismatch at the second nucleotide. To cleave either the on-target (grey, dashed line) and the off-target (blue, solid line), the largest energetic penalty comes from making it passed the ‘seed’ (nt 8). The off-target has raised this barrier (from the grey horizontal line to the blue horizontal line) by an amount equal to the second mismatch penalty seen in **Figure ??**C. The increased barrier exponentially suppresses the corresponding off-target (effective) cleavage rate. Given cleavage implies binding, also the effective association rate is exponentially suppressed. Placing the mismatch further down the hybrid, for example at nt 10, we see both binding and cleavage rates have recovered partially from their values in the seed (**Figure ??**A). The corresponding landscape in **Figure ??**C shows that the seed still imposes the largest barrier against cleavage, and thereby also against binding. Although raising the energy, and the barrier against R-loop completion, the energy for the off-target, also after nt 10, remains almost at the same height as the on-target landscape’s height in seed (compare the grey and blue horizontal lines). In other words, the mismatch therefore only minimally raises the effective barrier opposing R-loop completion. Hence, both dCas9 and Cas9 can complete R-loop formation at rates closer to that of completing the R-loop for the on-target.

Interestingly, placing a mismatch between nt 12–17 significantly reduces the cleavage rate, while only minimally impacting the association rate (**Figure ??**A). **Figure ??**D, displaying a landscape with a mismatch at nt 15, reveals that although binding (making it into any long-lived bound state) is limited mainly by the seed, cleavage necessitates proceeding past the second large barrier – now of similar height – seen beyond the 13th base pair. Hence, (d)Cas9 will bind such a target at a rate comparable to the on-target and get trapped in a configuration with a partial R-loop (the ‘intermediate state’ referred to above, **Figures ??**A-B). Eventually, Cas9 escapes from this intermediate, either through unbinding or cleavage, both requiring it to overcome a second large energetic barrier, thereby leading to relatively low cleavage rates at such off-targets, diverging from the relative association rate.

Besides providing Cas9 the ability to swiftly reject off-targets without matching seeds, the associated energetic barrier (between the ‘open’ and ‘intermediate’ configurations discussed above) significantly opposes cleavage of even the on-target. Raising this barrier further as a means to gain specificity, definitely reduces the efficiency at which Cas9 cleaves the on-target. The introduction of the second barrier separating the intermediate and closed states in the on-target free-energy landscape (**Figures ??**A and **??**B) allows Cas9 to reject an additional set of off-targets, without having to sacrifice the rate at which it can cut the on-target – preventing the first barrier from becoming of insurmountable height. Therefore, the promiscuous binding of Cas9 can be seen as a price to pay in order to be both a (sufficiently) fast and specific nuclease.

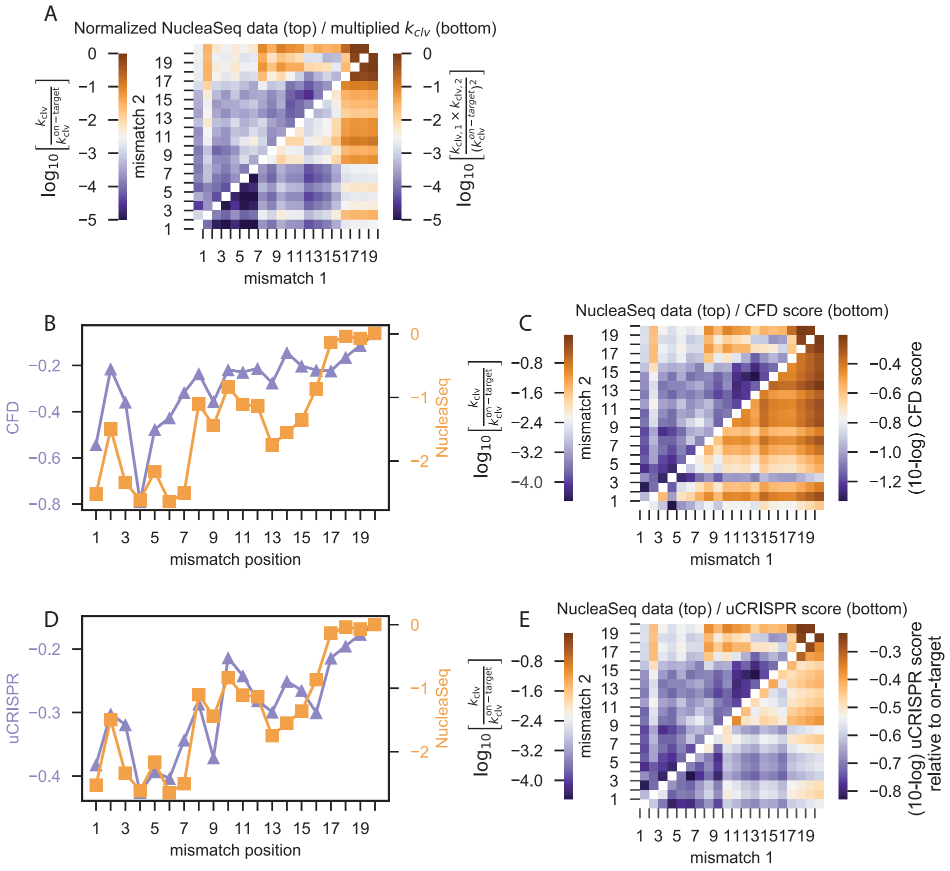


Figure 3.5: Comparison to thermodynamics based models. **(A)** Upper half shows NucleaSeq data for double mismatched off-targets, normalized to the on-target's rate. Bottom half uses single-mismatch data from **Figure ??**A as a naïve Bayes classifier to predict the double-mismatch data. For every set of two mismatch positions, the lower half shows the product of the corresponding data points from **Figure ??**A. **(B)** Sequenced averaged CFD score compared to NucleaSeq data for off-targets with one mismatch (**Figure ??**A). **(C)** Sequenced averaged CFD score compared to NucleaSeq data for off-targets with two-mismatches (upper half **Figure ??**B) **(D)** Sequenced averaged uCRISPR score (normalized to on-target) compared to NucleaSeq data for off-targets with one mismatch. **(E)** Sequenced averaged uCRISPR score (normalized to on-target) compared to NucleaSeq data for off-targets with two-mismatches.

3.2.6. Existing off-target prediction models can be seen as a limiting case of ours

Currently, state-of-the-art off-target prediction [?] is based mainly on the 'Cutting Frequency Determination' (CFD) score [?] – a 'naïve Bayes classification' scheme [?] assuming mismatches affect the relative cleavage rate independent of the distance between them. More recently, Zhang et al. report their 'unified CRISPR' (uCRISPR) score [?], in which cleavage probabilities are evaluated as the Boltzmann weight corresponding to the cleavage competent state (see **S.I.**), outperforms the CFD score.

In the **S.I.** we show that both models can be seen as a limiting case of ours. To reduce our model to theirs, we must assume target-binding equilibrates prior to cleavage and that all bound states are unstable compared to solution (independent of nuclease concentration) (see **S.I.**). Within this limit, the relative rate of cleaving a multi-mismatched off-target versus the on-target equals the product of the corresponding relative rates of cleaving the set of singly mismatched off-targets (see S.I. for details). For example, if an off-target has mismatches at positions 5 and 7, its corresponding relative rate equals the product of the relative rates for cleaving the off-targets with one mismatch at nt 5 and the one with a mismatch at nt 7, all compared to the on-target cleavage rate. As this is exactly how the CFD score has been constructed using their own set of experiments [**?**], a special case of the mechanistic model presented here produces a score equal to the CFD score – despite the construction of the CFD score not being motivated by physics. Furthermore, our model directly reduces to the uCRISPR score within these same limits (**S.I.**).

The physical regime wherein CFD and uCRISPR could ever produce accurate predictions corresponds to all bound states, including the cleavage competent state being energetically unfavorable compared to solution, no matter the nuclease concentration. This regime clearly does not comply with free-energy estimates, even at 1nM (d)Cas9-sgRNA (**Figure ??**A). We take the quantitative agreement between our model and the bulk experimental data (**Figure ??**), and its consistency with single-molecule data (**Figure ??**), to imply the physical regime suggested by our model parameters to be valid.

As assuming no cooperative effect of mismatches (as done in by the mentioned equilibrium based models) is an attractive approach due to its simplicity, it is informative to see exactly where it fails. To test whether a naïve Bayes classifier can be used as an accurate predictor of the NucleaSeq data for the given sgRNA, we first test whether products of relative k_{clv} values for singly mismatched off-targets in the NucleaSeq dataset are a good predictor of the corresponding measurements at off-targets containing two mismatches (**Figure ??**A). **Figure ??**A shows the NucleaSeq data normalized to the on-target cleavage rate. While the upper half displays the normalized data directly, the bottom half is constructed by using products of the measured single-mismatch values (**Figure ??**A). Clearly, assuming no cooperative effect of mismatches does not result in the measured (relative) cleavage rates. In particular, the cleavage rate is severely overestimated when both mismatches are placed outside the seed (beyond nt 8), but before nt 16. That is, when the mismatches are placed in between the 'intermediate' and 'closed' states (**Figure ??**B), which is exactly the set of off-targets that tend to lead to a divergence between apparent binding and cleavage rates (**Figure ??**).

Next, directly comparing the CFD score (**Figures ??**B,C) and the uCRISPR score (**Figures ??**D,E) to the (normalized) Nucleaseq data, we see both methods seem to be plagued by this same underestimation due to the non-additive nature of mismatches. The CFD score completely fails to produce even qualitatively similar relative rates (**Figures ??**B,C). Note that the method used in **Figure ??**A represents an equivalent CFD score, had the authors' used the Nucleaseq assay to produce their data, thereby showing it is the underlying assumptions of the CFD score rather than the training data that leads to inaccurate predictions. Furthermore, the uCRISPR score produces a single-mismatch profile similar to the Nucleaseq data (**Figure ??**D). Therefore, as mismatches also act independently within uCRISPR (see **S.I.**), this leads to a two-mismatch profile nearly identical to the one shown in **Figure ??**A, despite the large difference in absolute values compared to the data. Hence, even though the uCRISPR model has introduced an additional energetic penalty for placing consecutive mismatches (see supplement of [**?**]), it still ranks off-targets almost identical to a model that assumes mismatches each effect the cleavage rate independently.

Taken together, the more general kinetic model presented in this work correctly treats how multiple mismatches alter cleavage rates, and how binding does not imply cleavage, while the equilibrium based CFD and uCRISPR fail to do such.

3.2.7. Measuring relative rates at various concentrations

Thus far we have presented a physical model capable of explaining experimental data of various forms (**Figures ??** and **??**), and demonstrated the added benefit of incorporating the kinetics of the targeting process (**Figures ??** and **??**). In what remains, we shall use our model to predict cleavage rates under various experimental conditions.

Figures ??A and B show cleavage rates, normalized to on-target values, for several Cas9-sgRNA concentrations. First, we note that as the concentration is decreased, the ratio of cleavage rates (symbols in **Figure ??**A) approaches the ratio in probabilities for a (PAM) bound Cas9-sgRNA to cleave the DNA prior to rejecting it (pink line). This cleavage probability is the central quantity of **Chapter ??** [**?**] and we here confirm its validity in the low concentration regime.

Interestingly, varying the concentration mainly effects the relative cleavage rate at off-targets with PAM distal mismatches. **Figure ??** shows that by lowering the concentration the height of the effective barrier separating the open and intermediate states increases relative to the one separating intermediate and closed configurations. Hence, at low concentrations the contribution of this second transition to the cleavage rate is reduced, which manifests itself in an increase in the rates of cleaving correspondingly mismatched off-targets (a less sever 'dip' between positions 13 and 17) (**Figure ??**A). A similar signature is seen when comparing mismatches with two mismatches at 0.01nM and 100nM Cas9-sgRNA (**Figure ??**B). Lowering the concentration causes the effective cleavage rate to become limited by the rate of binding a DNA sequence from solution, multiplied by the probability to cleave once bound ($k_{\text{clv}} \approx k_{\text{on}} P_{\text{clv}}$, as k_{on} becomes rate limiting at low concentrations, **Figure ??** and **Chapter ??**).

3.2.8. Measuring relative fractions of cut DNA after various incubation times

Other than the concentration, the exposure time of the DNA to Cas9-sgRNA can be varied experimentally. **Figures ??**C and D show the relative probability of cleaving off-targets (compared to on-target) for different incubation times. When considering off-targets with a single mismatch, placing a mismatch directly adjacent to the PAM results in the lowest cleavage rate. If the experiment runs for a time exceeding the inverse of this rate (the

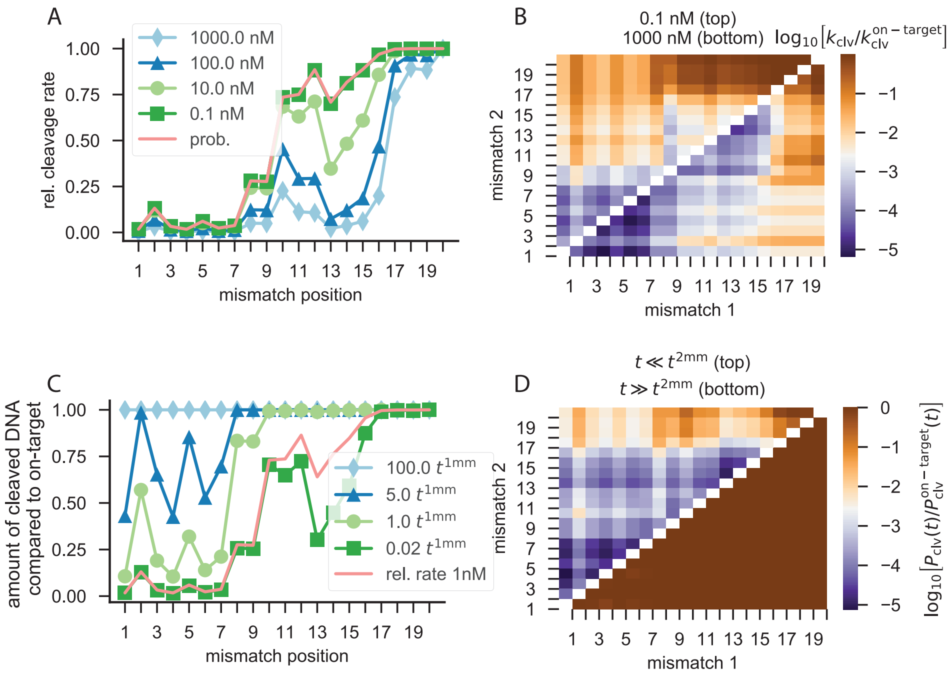


Figure 3.6: Measuring cleavage activity under varying experimental conditions. **(A)** Cleavage rates, normalized to on-target, for various nuclease concentrations (symbols). Solid line (pink) shows the probability that a PAM bound Cas9-sgRNA cuts the DNA before unbinding (relative to on-target) (**Equation ??**). **(B)** Relative cleavage rates for 0.01nM Cas9-sgRNA (upper half) and 100nM (lower half) Cas9-sgRNA. **(C)** Probability of a DNA target being cut, relative to on-target, after a fixed time (different symbols) and 1nM Cas9-sgRNA. $t^{1\text{mm}}$ represents k_{clv}^{-1} for the off-target with a mismatch adjacent to the PAM, which is the off-target with the lowest cleavage rate amongst all off-targets with a single mismatch. Solid link (pink) shows ratio between cleavage rates (off-target vs. on-target). **(D)** $t^{2\text{mm}}$ represents k_{clv}^{-1} for the off-target with mismatches at the first two positions adjacent to the PAM, which is the off-target with the lowest cleavage rate amongst all off-targets with two mismatches. Upper half $t = 10^{-5}t^{2\text{mm}} = 0.1t^{\text{on-target}}$, lower half shows $t = 100t^{2\text{mm}}$.

maximum k_{clv}^{-1} denoted by $t^{1\text{mm}}$ in **Figure ??**C), essentially any off-target (with a single) mismatch will get cut. Hence, no difference between off-targets and on-target will be observed when counting the relative fractions of cleaved DNA (light blue diamonds in **Figure ??**C). Similarly, using $t^{2\text{mm}}$ to denote the inverse cleavage rate for the off-target with mismatches at the first two R-loop nucleotides, all measured cleavage rates approach the off-target rates for incubation times exceeding $t^{2\text{mm}}$ (**Figure ??**D). Performing this same experiment after much shorter incubation times (dark green squares), we see that for mismatches in the seed, these relative counts are well approximated by the relative cleavage rates at the corresponding nuclease concentration (pink line in **Figure ??**C or the curve for 1 nM in **??**A). In the **S.I.** we show this implies the cleavage probability is well approximated by a single-exponential process. Placing the mismatch between intermediate and closed states increases the time to surpass the intervening barrier. When the time to transition into the closed state becomes comparable to the time to transition into the intermediate state from PAM, we expect the probability to cleave a DNA not anymore to follow a single exponential curve. For this reason the ratio in the cleavage rates does not anymore match the ratio in counted cleaved molecules when a mismatch is placed between positions 11 and 16 (**Figure ??**C). In conclusion, the incubation time greatly influences the relative fractions of cut DNA, both for PAM proximal as well as PAM distal mismatches (**Figures ??**C and D).

3.3. Discussion

The increasing popularity of the CRISPR-Cas9 system as a genome-editing tool calls for a quantitative understanding of its risks. Here, we presented a single mechanistic model (**Figure ??**A) to describe the kinetics of off-targeting by Cas9-sgRNA, as well as binding by the nuclease inactive dCas9-sgRNA. Using a (target) sequence averaged approach, we demonstrated our model accurately describes experimental association rates (**Figure ??**B), cleavage rates (**Figure ??**C) and dissociation constants (**Figure ??**D). The free-energy landscape(s) describing interactions between (d)Cas9-sgRNA with on-target (**Figures ??**A and D) and off-target DNA (**Figures ??**B-D) serve as our model parameters. Hence, using the bulk data (**Figure ??**B and C), we extracted the microscopic thermodynamic and kinetic properties of Cas9-sgRNA (**Figure ??**). The particular free-energy landscape obtained shows signatures consistent with Cas9's major conformational change, rearrangement of its nuclease domains, observed in structural and single-molecule experiments (**Figure ??**). Moreover, the barriers opposing this conformational change directly explains how Cas9's promiscuity when it comes to off-target binding is the price to pay for it to balance on-target and off-target cleavage activities (**Figure ??**). Further, the free-energy landscape implies Cas9 operates far from the regime in which existing prediction models operate. As a result, only our model quantitatively describes the difference between Cas9 and dCas9 specificities (**Figure ??**). Finally, we showed how varying nuclease concentrations and incubation times strongly influence, not only the quantitative, but also the qualitative specificity profiles (**Figure ??**).

3.3.1. Comment on translation to other guide RNA sequences ('shortcut' to redoing measurement for every guide)

In **Figures ??**A-C, we display target sequence averaged cleavage activities (w.r.t on-target) from datasets across the literature [**? ? ?**], including the data used to construct the CFD score (**Figures ??**B,C). Different curves correspond to different guide sequences. Also, **Figure ??**D shows a second NucleaSeq dataset (together with the data shown in **Figure ??**C) [**?**]. Clearly, the cleavage rate is strongly dependent on the guide sequence used.

As a future improvement to our model parameterization, incorporating (guide) sequence dependencies seems the most logical way forward. However re-training our model against equivalent datasets (HiTS-FLIP + NucleaSeq) [**? ?**] for every guide sequence of interest would require an immense amount of experimental effort.

For this reason, developing a translation between guides, using the current parameter set could be an attractive approach. **Figure ??** showed our model is capable of producing a wide range of specificity profiles by varying the experimental conditions. This variation appears to be similar to that caused by the guide sequence shown in **Figure ??**. For example, **Figure ??**A shows data belonging to six guides from Doench et al. [**?**]. A translation from the lower three curves (pink, orange, yellow) to the upper three curves (purple, green, blue) seems to be achievable within our model by a combination of lowering the concentration (**Figure ??**A) and increasing the incubation time (**Figure ??**C). This similarity between **Figures ??** and **??** leads us to believe a less experimentally (as well as computationally) intensive scheme may exist to predict off-targeting for different guides. Differences in sequence manifest themselves in the energetics (**Figures ??**A-C), altering (effective) barrier heights separating states. **Figure ??** demonstrated the same can be achieved through varying the Cas9-sgRNA concentration. Alternatively, increasing the incubation time increases the probability of exceeding the typical time needed to reach states further down the landscape (**Figure ??**A). We hypothesize that varying the guide sequence could possibly be modeled by an altered 'effective nuclease concentration' and 'effective experimental time', while keeping the same model parameters (**Figure ??**) as determined for the guide used in this work. In this manner, sequence dependencies can possibly be derived from experiments performed for a limited set of guides.

3.3.2. Move to other guided nucleases (generality of approach)

Cas9-sgRNA is by far not the only RNA guided nuclease system utilized in biotechnological applications. Other CRISPR associated nucleases, such as CRISPR-Cas12a, CRISPR-Cas13 and CRISPR-Cas14 offer a diversified 'genome-engineering toolkit' to complement Cas9 [**? ? ? ? ? ? ?**]. Moreover, the high-specificity requirements for therapeutic applications has driven the development of several strategies to improve Cas9's cleavage specificity, with the use of either engineered [**? ? ?**] or natural variants (such as *N. meningitides* Cas9) [**?**] becoming increasingly popular. The general model presented in this work (**Figure ??**A) should be applicable to any RNA guided nuclease whose target binding happens in a sequential fashion. High-throughput measurements using different nuclease systems (preferably similar to HiTS-FLIP, CHAMP and/or NucleaSeq, i.e. [**? ? ? ?**]), will allow us to also decipher their microscopic free-energy landscape underlying target interference and can point towards the relevant structure-function relations (as done here for *Streptococcus pyogenes* Cas9).

3.3.3. Test against genome-wide off-target data/prediction tools will follow

Existing off-target prediction tools [**? ? ? ? ? ? ? ?**] are not all made to quantitatively predict experimental measurements, but rather to rank off-targets according to their activity (w.r.t on-target). Typically, the performance is assessed using either of two methods. Either the rank correlation between modeled scores and measurements is used as a performance measure [**? ?**]. Alternatively, prediction tools are tested for their capability to separate the 'cut' from 'uncut' genomic DNA sites [**?**]. Although our physical model offers more than such a classification scheme, we nevertheless are working towards performing tests against identified genomic off-targets [**? ?**] in order to directly compare our model to other bioinformatics or machine learning based predictors.

3.4. Acknowledgments

Special thanks to excellent work done by Stijn van der Smagt and Koen van der Sanden (both Msc Applied Physics). Koen helped during the early stages of this project in developing the fitting code. Stijn performed the combined fit using the HiTS-FLIP and NucleaSeq data. This work was supported by the Netherlands Organization for Scientific Research (NWO/OCW), as part of the Frontiers in Nanoscience program.

3.5. Supplemental Information

3.5.1. Kinetic model for Target Recognition

We here explain in more precise mathematical terms how we have built the model put forth in the main text and in **Figure ??**A. To incorporate the concentration of Cas9-sgRNA present in solution ('sol'), we shall take the viewpoint of a single DNA target sequence, either on- or off-target. After one of the Cas9-sgRNA binds the DNA at its PAM site, R-loop formation (Cas9 mediated strand exchange between gRNA and DNA) is modeled as a sequential process. That is, the gRNA-DNA hybrid grows or shrinks with single-nucleotide increments, allowing for hybrids of intermediate lengths (1-20 bp formed). Cleavage ('clv') can follow complete R-loop formation (20 nucleotides in case of Cas9). Together, we model the entire target recognition process as a random walk on the linear state-space, $n \in [\text{sol}, \text{PAM}, 1, 2, ..., 20, \text{clv}]$. Knowing the probability of a Cas9-sgRNA-DNA to be found in each of the states after a time t gives access to any measurable quantity of interest (see below for examples). Letting $P_{\text{n}}(t)$ denote the occupancy of state n at time t, and $k_{\text{f}}(n)/k_{\text{b}}(n)$ the rates (inverse average times) for 'forward'($n \rightarrow n+1$)/'backward'($n \rightarrow n-1$) transitions, the occupancies evolve according to the following set of Master Equations

$$\frac{\partial P_{\text{sol}}}{\partial t} = -k_{\text{f}}(\text{sol})P_{\text{sol}}(t) + k_{\text{b}}(\text{PAM})P_{\text{PAM}}(t) \tag{S3.1}$$

$$\begin{aligned}\frac{\partial P_{\text{n}}}{\partial t} &= k_{\text{f}}(n-1)P_{\text{n-1}}(t) - (k_{\text{f}}(n) + k_{\text{b}}(n))P_{\text{n}}(t) \\ &+ k_{\text{b}}(n+1)P_{\text{n+1}}(t) \qquad \forall n \in [\text{PAM}, 1, 2, ..., 19]\end{aligned} \tag{S3.2}$$

$$\frac{\partial P_{20}}{\partial t} = k_{\text{f}}(19)P_{19}(t) - (k_{\text{f}}(20) + k_{\text{b}}(20))P_{20}(t) \tag{S3.3}$$

From here on we interchangeably use $n = -1 \equiv$ sol, $n = 0 \equiv$ PAM and $n = 21 \equiv$ clv. Given any DNA is either unbound, bound or cleaved, the fraction of cleaved DNA (for active Cas9) is set by $P_{\text{clv}}(t) = 1 - \sum_{n \neq \text{clv}} P_{\text{n}}(t)$. Defining the vector $\vec{P}(t) = [P_{\text{sol}}(t), P_{\text{PAM}}(t), P_1(t),....,P_{20}(t)]^T$, the solution to **Equations ??** and **??** can be written as

$$\vec{P}(t) = e^{-Kt}\vec{P}(0), \tag{S3.4}$$

with the (tri-diagonal) rate matrix K's elements given by

$$K_{n,m} = \begin{cases} -k_{\text{f}}(n-1) & n = m+1 \\ k_{\text{f}}(n) + k_{\text{b}}(n) & n = m \\ -k_{\text{b}}(n+1) & n = m-1 \\ 0 & \text{else} \end{cases} \tag{S3.5}$$

In general, we recognize the system is completely determined by the set of forward and backward rates for every Cas9-sgRNA-DNA of interest. To extract information from the experimental data, we now proceed to show the particular parameterization used throughout this work.

As mentioned in the main text, we have chosen a DNA sequence averaged parameterization. Adding any nucleotide to the R-loop is assumed to happen at the same rate (denoted by k_{f}, as opposed to the general position dependent forward rate $k_{\mathrm{f}}(n)$). Further, the rate of binding from solution onto the DNA (transitioning from sol ($n = -1$) to PAM ($n = 0$)) is assumed to grow linearly with concentration, $k_{\mathrm{on}} = k_{\mathrm{on}}([\text{Cas9-sgRNA}]_{\mathrm{ref}}) \times [\text{Cas9-sgRNA}]$, resulting in the binding rate at our chosen reference concentration $[\text{Cas9-sgRNA}]_{\mathrm{ref}} = 1\text{nM}$ being a free-parameter. Finally, catalyzing the reaction to induce the DNA breaks is assigned a separate rate of k_{cat}. Taken together, forward transitions are assigned the following rates

$$k_{\mathrm{f}}(n) = \begin{cases} k_{\mathrm{on}}([\text{Cas9-sgRNA}]) & n = \text{PAM} \\ k_{\mathrm{f}} & n \in [1, 2, ..., 19] \\ k_{\mathrm{cat}} & n = 20 \end{cases} \tag{S3.6}$$

Backward rates (unbinding, shrinking the R-loop) are set by requiring the convergence of $P_{\mathrm{n}}(t)$ to the Boltzmann Distribution when equilibrated.

$$P_{\mathrm{n}}^{\mathrm{EQ}} = \frac{e^{-F_{\mathrm{n}}}}{\sum\limits_{m \in [\text{sol,PAM},1...,20]} e^{-F_{\mathrm{m}}}} \quad \forall n \in [\text{sol}, \text{PAM}, 1..., 20] \tag{S3.7}$$

Given all occupancies are time-independent in this limit ($\frac{\partial \vec{P}}{\partial t} = 0$), **Equations ??-??** result in the 'detailed balance condition'

$$k_{\mathrm{b}}(n) = k_{\mathrm{f}}(n-1)\frac{P_{\text{n-1}}^{\mathrm{EQ}}}{P_{\mathrm{n}}^{\mathrm{EQ}}} = k_{\mathrm{f}}(n-1)e^{F_{\mathrm{n}} - F_{\text{n-1}}} \quad \forall n \in [\text{PAM}, 1, ..., 20] \tag{S3.8}$$

Differences in free-energy (F_{n}'s, measured in units of $k_B T$) between consecutive states for a particular Cas9-sgRNA-DNA are modeled as ($F_{\mathrm{sol}} = 0$ as reference state)

$$F_{\mathrm{n}} - F_{\text{n-1}} = \begin{cases} F_{\mathrm{PAM}}([\text{Cas9-sgRNA}]) & n = \text{PAM} \\ \epsilon_C(n) & \text{match at } n \in [1, 2, ...20] \\ \epsilon_C(n) + \epsilon_I(n) & \text{mismatch at } n \in [1, 2, ...20] \end{cases} \tag{S3.9}$$

If the n$^{\text{th}}$ base of the target is complementary to the corresponding base of the guide, the Cas9-sgRNA-DNA ternary complex gains/loses $\epsilon_C(n)\ k_B T$ in incorporating the basepair into the R-loop. The Cas9 protein is known to interact with the (non-target strand) DNA, as well as undergo conformational changes, during the process of R-loop formation. For this reason, $\epsilon_C(n)$'s can either be negative (signifying an energetic benefit) or positive (penalizing progression of the R-loop). If the n$^{\text{th}}$ base of the target does not match the guide's base, the ternary complex gets penalized $\epsilon_I(n) \geq 0$ for incorporating the mismatch into the R-loop. All subsequent free-energy states are therefore also raised by this same amount

(**Figure ??**B), thereby only locally increasing the backward rate (**Equation ??**).
The energy of the PAM bound configuration is modeled as a concentration dependent free-energy, $F_{\text{PAM}}([\text{Cas9-sgRNA}]) = F_{\text{PAM}}([\text{Cas9-sgRNA}]_{\text{ref}}) - \log([\text{Cas9-sgRNA}])$, becoming more stable with increasing nuclease concentration. Note that using the concentration dependencies of both F_{PAM} and k_{on}, via **Equation ??**, leads to a concentration independent rate to return to solution $k_{\text{b}}(\text{PAM})$.

In conclusion, we have built a general kinetic model (**Equation ??**), and used a DNA sequence averaged parameterization to reduce our parameter space to the following 44 parameters: (1) $F_{\text{PAM}}([\text{Cas9-sgRNA}]_{\text{ref}})$, (2-21) 20x $\epsilon_C(n)$'s, (21-41) 20x $\epsilon_I(n)$'s,
(42) $k_{\text{on}}([\text{Cas9-sgRNA}]_{\text{ref}})$, (43) k_{f}, and (44) k_{cat}. When considering dCas9 cleavage is unable to occur, which is simply modeled by setting $k_{\text{cat}} = 0$ (leaving 43 free-parameters).

3.5.2. Calculating (effective) association rates (HiTS-FLIP)

To predict measured association rates, we assume equivalence between the solution to the Master Equations (**Equation ??**) and the fluorescence signal obtained in the HiTS-FLIP experiment [**?**]. Experiments are performed at 1nM dCas9-sgRNA, which we thereby set as our reference concentration. Given the experiment uses dCas9, all molecules are either in solution or bound to DNA ($P_{\text{clv}} = 0$). Here we follow the procedure detailed in Boyle et al. [**?**]. First, we determine the fraction of bound DNA molecules,

$$P_{\text{bnd}}(t) = \sum_{n \in \{\text{PAM},1,..20\}} P_{\text{n}}(t) = 1 - P_{\text{sol}}(t) \tag{S3.10}$$

at three specified time points $t_1 = 500s$, $t_2 = 1000s$ and $t_3 = 1500s$, starting with all DNA molecules being unbound at $t_0 = 0s$ ($P_{\text{sol}}(0) = 1$, $P_{\text{n}}(0) = 0 \; \forall n \neq \text{sol}$). Next, the effective association rate (k_{a}) is defined as the coefficient of a linear fit to the three occupancies, forced to go through the origin,

$$p_i = k_{\text{a}} t_i \;\; \forall i \in [0,1,2,3] \tag{S3.11}$$

Equation ?? is the approximate solution for $P_{\text{bnd}}(t)$ for $t \ll k_{\text{a}}^{-1}$, if one would assume the system not to consist of 21 possible bound states (as done here), but just by a single one. Namely, in this simplified two-state system ($n \in [\text{sol}, \text{bnd}]$)

$$\frac{\partial P_{\text{bnd}}}{\partial t} = k_{\text{a}} P_{\text{sol}}(t) \Rightarrow P_{\text{bnd}}(t) = 1 - e^{-k_{\text{a}} t} \approx k_{\text{a}} t \;\text{ if } t \ll k_{\text{a}}^{-1} \tag{S3.12}$$

Using least-squares optimization (linear regression),

$$k_{\text{a}} = \frac{\sum_{i=1}^{3} t_i p_i}{\sum_{j=1}^{3} t_j^2} \tag{S3.13}$$

3.5.3. Calculating (effective) cleavage rates (NucleaSeq)

Next, we show how we instead can use the solution to the Master Equations (**Equation ??**) to mimic the NucleaSeq experiment [**?**]. NucleaSeq is performed at saturating concentrations of Cas9-sgRNA, which we model by setting $F_{\text{PAM}} \ll 0k_BT$ (we chose $F_{\text{PAM}} = -1000k_BT$), $k_{\text{on}} \to \infty$. As done in the original experiment, the fraction of DNA not cleaved,

$$P_{\text{no clv}}(t) = 1 - P_{\text{clv}}(t) = \sum_{n \neq \text{clv}} P_{\text{n}}(t) \tag{S3.14}$$

is evaluated at the time points t_0 through t_9 as 0 s,12 s,60 s,180 s,600 s,1800 s,6000 s,18000 s, and 60000 s (using the initial condition of everything being unbound at $t_0 = 0s$, which due to the high nuclease concentration results in (near) instantaneous occupation of the DNA). Similarly to Boyle et al., Jones et al. assume the system to consist of just a single bound state, for which the fraction of cleaved DNA under saturating conditions (no unbound DNA) follows

$$\frac{\partial P_{\text{clv}}}{\partial t} = k_{\text{clv}} P_{\text{no clv}}(t) \Rightarrow P_{\text{no clv}}(t) = e^{-k_{\text{clv}}t} \tag{S3.15}$$

Hence, we obtain the effective cleavage rate ($k_{\text{clv}}(t)$) by fitting a line (forced through origin) to the logarithm of the occupancies,

$$\log(p_i) = -k_{\text{clv}} t_i \; \forall i \in [0, 1, 2, ..., 9], \tag{S3.16}$$

Using linear regression,

$$k_{\text{clv}} = -1 \times \frac{\sum_{i=1}^{9} t_i \log(p_i)}{\sum_{j=1}^{9} t_j^2} \tag{S3.17}$$

3.5.4. Calculating apparent binding affinities (CHAMP)

A third quantity used throughout this work are 'Apparent Binding Affinities' (ABA) obtained from the CHAMP experiment [**?**]. CHAMP experiments are performed using dCas9, at varying nuclease concentrations, rather than varying incubation times. Using the fitted binding rate at 1nM,

$$k_{\text{on}}([\text{Cas9-sgRNA}]) = k_{\text{on}}^{\text{1nM}} \frac{[\text{Cas9-sgRNA}]}{[1 \text{ nM}]} \tag{S3.18}$$

The experiment consists of determining the fraction of bound DNA, **Equation ??**, at $t = 10$ minutes, for the concentrations ([Cas9-sgRNA]) 0.1 nM,0.3 nM,1 nM,3 nM,10 nM,30 nM,100 nM and 300 nM. Assuming the system has had sufficient time to equilibrate within these 10 minutes, the series of occupancies should follow the Hill Equation (using $c = {}^{[\text{Cas9-sgRNA}]}/_{[1\text{nM}]}$ to denote the relative concentration)

$$H = \frac{1}{1 + \frac{K_D}{c}} \tag{S3.19}$$

A fit of **Equation ??** to the series of occupancies (for the specified concentrations), results in the apparent half-saturation concentration, or apparent dissociation constant K_D for the (off-)target of interest. The ABA is defined as the logarithm of K_D, which has units of free-energy. The quantity shown are what are termed, ΔABA's, which are ABA differences w.r.t. the on-target

$$\Delta ABA = ABA_{\text{off-target}} - ABA_{\text{on-target}} \tag{S3.20}$$

3.5.5. Simulated Annealing fitting

All fits are performed using a custom written Simulated Annealing (SA) algorithm to minimize the χ^2 (least-squares optimization). Prior to fitting, the data (k_{a} and k_{clv} values) are converted to sequence averaged values for every unique mismatch pattern, weighted by the square of their corresponding measurement errors (σ(sequence)),

$$k_{\text{clv/a}}(\text{mm pattern}) \equiv \sum_{\text{the same mm-pattern}:i} w_i k_{\text{clv/a},i}\,, \quad w_i = {}^{\sigma_i}/_{\sum_j \sigma_j} \tag{S3.21}$$

The sums in **Equation ??** run over all off-target sequences in the library that have the same mismatch pattern. This particular weighted average is chosen as one can prove that it represent the best possible sequence averaged model - it is the global optimal χ^2 when allowing one to assign exactly one model value to every possible mismatch pattern. Hence, a good fit to the weighted averaged data represents a good fit to the raw data. The corresponding measurement error in the weighted averaged rates ('standard error in the weighted mean') follows

$$\hat{\sigma}(\text{mm pattern}) \equiv \sum_{\text{the same mm-pattern}:i} w_i^2 \sigma_i^2, \tag{S3.22}$$

Furthermore, in our experience we obtained more accurate predictions of the lower valued k_{clv}'s in the NucleaSeq experiment when fitting not to the k_{clv}'s, but to the $\log(k_{\text{clv}})$ values in stead. For consistency we therefore also fitted against $\log(k_{\text{a}})$ values (in case of the simultaneous fit). To construct a global χ^2 for both association and cleavage rate experiments, the individual χ^2's are added together after dividing each by the number of different sequences with the identical mismatch pattern in the respective libraries. For both libraries, each member sequence contains more than the 20 nucleotides + 3 PAM nucleotides that are important for targeting. Hence, multiple members would be considered to be an on-target (also because of the first nucleotide in the NGG PAM that is allowed to vary). Similarly, more than 3 off-targets are present with a single mismatch at one of the 20 R-loop positions. Using i to iterate over unique mismatch patterns, we let N_i denote the number of library members with pattern i. Further, their simply are more unique mismatch patterns with two mismatches (${}^{20\times19}/_{2} = 190$ in total) than with a single mismatch (20 in total). To not over represent the influence of sequences with two mismatches, compared to single mismatches (and on-targets), the χ^2 is further divided into individual terms with fixed total number of mismatches, dividing by the total number of unique mismatch configurations in every group,

$$\begin{aligned}\chi^2 = &\frac{1}{N_{\text{on-target}}} \sum_{\text{on-targets}} \left(\frac{\log(k_{\text{a}}^{\text{model}}) - \log(k_{\text{a}}^{\text{experiment}})}{\hat{\sigma}} \right)^2 \\ &+ \frac{1}{20} \sum_{\text{single mm position:}i} \frac{1}{N_i} \left(\frac{\log(k_{\text{a,i}}^{\text{model}}) - \log(k_{\text{a,i}}^{\text{experiment}})}{\hat{\sigma}_i} \right)^2 \\ &+ \frac{1}{190} \sum_{\text{double mm pattern:}i} \frac{1}{N_i} \left(\frac{\log(k_{\text{a,i}}^{\text{model}}) - \log(k_{\text{a,i}}^{\text{experiment}})}{\hat{\sigma}_i} \right)^2 \\ &+ \frac{1}{N_{\text{on-target}}} \sum_{\text{on-targets}} \left(\frac{\log(k_{\text{clv}}^{\text{model}}) - \log(k_{\text{clv}}^{\text{experiment}})}{\hat{\sigma}} \right)^2 \\ &+ \frac{1}{20} \sum_{\text{single mm position:}i} \frac{1}{N_i} \left(\frac{\log(k_{\text{clv,i}}^{\text{model}}) - \log(k_{\text{clv,i}}^{\text{experiment}})}{\hat{\sigma}_i} \right)^2 \\ &+ \frac{1}{190} \sum_{\text{double mm pattern:}i} \frac{1}{N_i} \left(\frac{\log(k_{\text{clv,i}}^{\text{model}}) - \log(k_{\text{clv,i}}^{\text{experiment}})}{\hat{\sigma}_i} \right)^2\end{aligned} \tag{S3.23}$$

Here, $k^{\text{experiment}}$ and $\hat{\sigma}$ values are given by **Equations ??** and **??**. The model's values $k^{\text{experiment}}$ are determined using **Equations ??** and **??**.

The SA algorithm [**?**] is commonly used for high-dimensional optimization problems, such as the fit presented here, and we here highlight the specific adjustments made to suit our problem. In brief, the SA algorithm finds the (presumably) global minimum of the objective function $\chi^2(\vec{X})$, a function of the set of parameter values $\vec{X}$, by assuming equivalence to the potential energy of a physical system. In every iteration, the parameter vector is updated according to (letting $U(-\delta, \delta)$ denote the uniform distribution from $-\delta$ to δ)

$$\vec{X} \rightarrow \underbrace{\vec{X} + U(-\delta, \delta)}_{\vec{X}'} \tag{S3.24}$$

We shall refer to δ as the step size. After the update, the new parameter set ($\vec{X}'$) is accepted if it lowers the objective function ($\chi^2(\vec{X}') < \chi^2(\vec{X})$) or with a probability proportional to its corresponding Boltzmann weight when $\chi^2(\vec{X}') \geq \chi^2(\vec{X})$. The resulting 'acceptance probability' is known as the Metropolis condition,

$$p_{\text{acc}} = \min[1, \frac{e^{-\chi^2(\vec{X}')/T}}{e^{-\chi^2(\vec{X})/T}}] \tag{S3.25}$$

In the SA algorithm, the 'temperature' (T) is reduced iteratively to bias the system (parameter vector $\vec{X}$) to occupy its 'ground state' (global minimum of $\chi^2(\vec{X})$). We start from an initial temperature (T_0) as the temperature at which the initially supplied step size (δ)

results in an acceptance ratio between 40% and 60% (evaluated every 1000th iteration). Next, $\vec{X}$ is reset, and δ is adapted every 1000 iterations to ensure an acceptance ratio of 40-60% at the current temperature, before moving on to the next temperature after one more set of 1000 iterations. In analogy with statistical mechanics, we thereby let the system equilibrate at every temperature before moving onwards [**?**]. Here, we used an exponential cooling scheme with a 1% cooling rate for which the temperatures are defined by the series

$$T_k = 0.99^k T_0 \tag{S3.26}$$

The algorithm is stopped, minimum has been found, when both: (i) the temperature has fallen below 1% of its initial value $T < 0.01T_0$, and (ii) the relative change in average χ^2 (after equilibration), induced by reducing the temperature from T_k to T_{k+1}, has fallen below the user-defined threshold (10^{-5} for all reported fits)

$$\frac{|\langle\chi^2\rangle_k - \langle\chi^2\rangle_{k+1}|}{\langle\chi^2\rangle_k} \leq 10^{-5} \tag{S3.27}$$

with $\langle\chi^2\rangle_k$ denoting the average χ^2 at temperature T_k (determined in the 1000 steps after 'equilibration' as been reached, acceptance ratio of 40-60%). To be more confident that our presented solution represents the global optimum of χ^2, we repeat our SA fit several times, **Figures ??** and **??** presents the best solution amongst the different replica.

In **Figures ??**D-F we post-selected the final results from the individual runs of the algorithm by requiring that the resulting k_a values (the only quantity fitted in this figure) on average differ $\leq 5\%$ from those corresponding to the best fit. That is, the runs shown in **Figures S1D-F** satisfy

$$\frac{1}{\#\ \text{mm-patterns}} \sum_{\text{mm-pattern:}i} \frac{|k_{\mathrm{a,i}}^{\mathrm{run}} - k_{\mathrm{a,i}}^{\mathrm{best}}|}{k_{\mathrm{a,i}}^{\mathrm{best}}} \leq 0.05, \tag{S3.28}$$

which we take to be 'equally valid' solutions, as we now have filtered out fits clearly frozen into sub-optimal minima. For the simultaneous fit of **Figure ??**, no such selection was needed as all runs satisfied the equivalent of **Equation ??** with both k_a and k_clv.

3.5.6. Translation to models assuming individual mismatches act additively

Here we show in what limits our kinetic model corresponds to existing state-of-the-art (model-based) prediction tools, in particular CFD [**?**] and uCRISPR [**?**]. Although no direct comparison with our model has been given, we also discuss how the model of Farasat and Salis can be rationalized from ours [**?**]. Despite the different parameterizations, said models treat mismatches along the R-loop in quite similar fashion. To get the probability (relative rate) to cleave an off-target (compared to the on-target), the individual contributions of separate mismatches are either added together in energy-space (uCRISPR) or multiplied together in terms of their provabilities (CFD). From our physical model, we can understand what assumptions have (implicitly) been made in their construction, and therefore must hold in order to produce an accurate prediction.

As has been done explicitly when constructing uCRISPR, we start by assuming the PAM recognition and R-loop formation processes equilibrate prior to cleavage. In this limit, the effective rate of cleaving an off-target equals the fraction of Cas9-sgRNA-DNA that is in the cleavage competent state, multiplied by the bare catalytic rate,

$$k_{\text{clv}} \approx k_{\text{cat}} P_{20}^{\text{EQ}} \tag{S3.29}$$

When our sequential model equilibrates, occupancies follow the Boltzmann distribution,

$$P_{20}^{\text{EQ}} = \frac{e^{-F_{20}}}{1 + \sum\limits_{n=\text{PAM}}^{20} e^{-F_{\text{n}}}} \tag{S3.30}$$

The Boltzmann factor ($e^{-F_{20}}$) alone explains how straight addition of free-energy mismatch penalties lead to multiplication of probabilities. However, as seen in **Equation ??**, the Boltzmann factor merely describes the numerator and to calculate the probability one must first evaluate the partition function which is the denominator. Existing models used different versions of the partition function. First, Farasat and Salis, only account for the solution and cleavage competent states (state ’20’). Within the context of our microscopic model, this implies all but the final state’s energy are much greater than the solution state’s free-energy,

$$F_{\text{n}} \gg 1\ k_B T\ \forall n \in [\text{PAM}, 1...19] \Rightarrow P_{20}^{\text{EQ}} \approx \frac{e^{-F_{20}}}{1 + e^{-F_{20}}} \tag{S3.31}$$

This is the core of the model used by Farasat and Salis ([**?**]), in which F_{20} includes both sequence and position dependent mismatch penalties. In effect, both uCRISPR and CFD have further assumed also the cleavage competent state is unstable (compared to solution),

$$F_{20} \gg 1\ k_B T, \tag{S3.32}$$

which reduces the occupation to al but its corresponding Boltzmann weight

$$P_{20}^{\text{EQ}} \approx e^{-F_{20}}, \tag{S3.33}$$

The uCRISPR model uses **Equation ??** to determine (relative) cleavage rates (**Equation ??**), using a set of sequence and position dependent energies. To partially correct for their model’s inability of naturally explaining the non-additive nature multiple mismatches have, the authors used a set of additional energetic penalties for incorporating consecutive mismatches.

We note that **Equation ??** also describes the CFD model. CFD uses a set of measured probabilities to cleave singly mismatched off-targets w.r.t the on-target, which according to **Equation ??** amounts to measuring relative rates.

$$p \equiv \frac{P_{20}^{\text{EQ}}(\text{1x mm})}{P_{20}^{\text{EQ}}(\text{on-target})} \tag{S3.34}$$

The probability to cleave an off-target containing multiple mismatches, say at locations $mm1$ and $mm2$, is obtained by multiplying the individual probabilities for the off-targets containing either of the mismatches. To see that our general model also returns multiplications of probabilities in the limit where **Equation ??** is valid, we denote the free-energy of the off-targets (applying **Equation ??**)

$$F_{20}(\text{2x mm}) \equiv \epsilon_I(\text{mm 1}) + \epsilon_I(\text{mm 2}) + F_{20}(\text{on-target}) \tag{S3.35}$$

Next, using the approximations (**Equations ??** and **??**) leading up to **Equation ??**,

$$\begin{aligned}\frac{P_{20}^{\mathrm{EQ}}(\text{2x mm})}{P_{20}^{\mathrm{EQ}}(\text{on-target})} &= \frac{e^{-F_{20}(\text{2x mm})}}{e^{-F_{20}(\text{on-target})}} \\ &= e^{-(\epsilon_I(\text{mm 1})+\epsilon_I(\text{mm 2}))} \\ &= e^{-\epsilon_I(\text{mm 1})} \times e^{-\epsilon_I(\text{mm 2})} \\ &= \frac{e^{-F_{20}(\text{mm 1})}}{e^{-F_{20}(\text{on-target})}} \times \frac{e^{-F_{20}(\text{mm 2})}}{e^{-F_{20}(\text{on-target})}} \\ &= \frac{P_{20}^{\mathrm{EQ}}(\text{mm 1})}{P_{20}^{\mathrm{EQ}}(\text{on-target})} \times \frac{P_{20}^{\mathrm{EQ}}(\text{mm 2})}{P_{20}^{\mathrm{EQ}}(\text{on-target})} \\ &= p_1 \times p_2\end{aligned} \tag{S3.36}$$

Note that **Equation ??** represents the defining assumption of any 'naïve Bayes classifier' used to predict cleavage activities [**?**].

In conclusion, the models discussed here are only ever expected to produce accurate (relative) cleavage rates if any bound state is unstable, independent of Cas9-sgRNA concentration - an assumption that contradicts our model's parameterization (**Figure ??**).

3.5.7. At short times, relative counts equal relative rates

After exposing the DNA to Cas9-sgRNA for a time t, the number of DNA molecules cut equals the probability of any molecule being cleaved, $P_{\text{clv}}(t)$ given by **Equations ??** and **??**, multiplied by the total number of copies in the original pool of molecules (N_{pool}). Assuming the same copy number of every off-target tested in the experiment, letting $P_{\text{clv}}^{\text{on-target}}(t)$ denote the probability of a on-target DNA molecule being cleaved, the number of cleaved copies of an off-target compared to the number of cut on-targets equals

$$\frac{P_{\text{clv}}}{P_{\text{clv}}^{\mathrm{on-target}}} = \frac{1-e^{-k_{\text{clv}}t}}{1-e^{-k_{\text{clv}}^{\mathrm{on-target}}t}} \overset{t\to 0}{\approx} \frac{1-(1-k_{\text{clv}}t)}{1-(1-k_{\text{clv}}^{\mathrm{on-target}}t)} = \frac{k_{\text{clv}}}{k_{\text{clv}}^{\mathrm{on-target}}} \tag{S3.37}$$

, if the system can be approximated by the simpler **Equation ??**. We thus see that for short experiments, the fraction of cut DNA molecules can approach the fraction to the corresponding effective cleavage rates.

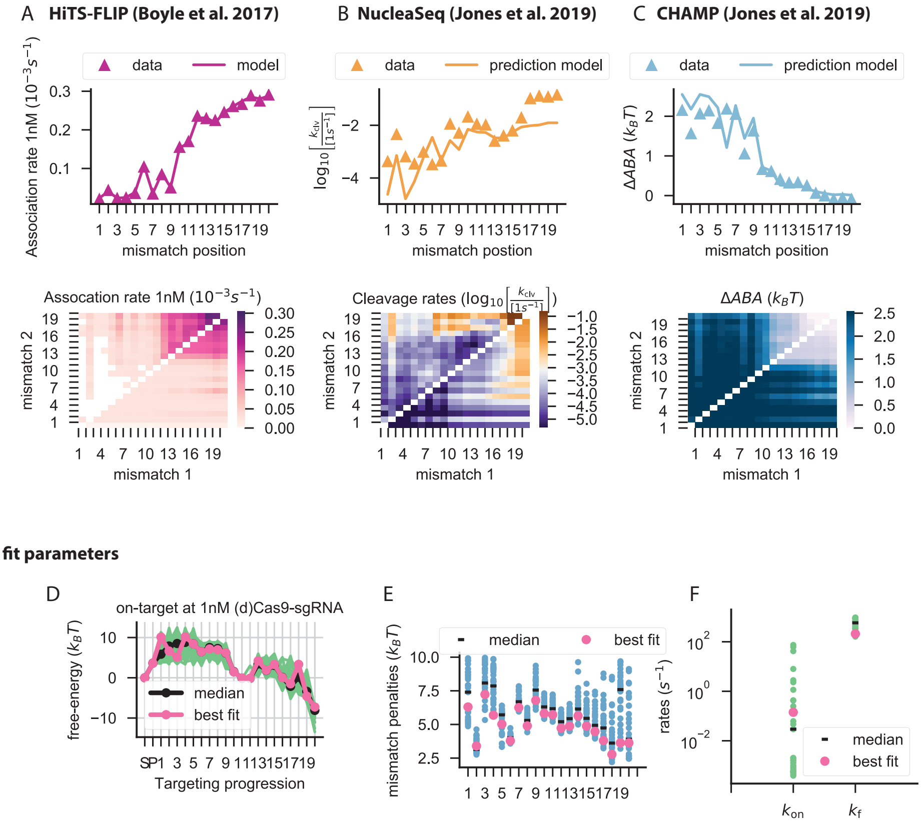


Figure S3.1: related to Figure ??. Fit only to dCas9 data from HiTS-FLIP (Boyle et al.) (A) Comparison of model to HiTS-FLIP data. top: fit against off-targets with 1 mismatch. bottom: fit against off-targets with 2 mismatches (data in upper triangle/ model in lower triangle). **(B)** Comparison of model to CHAMP data. top: prediction of off-targets with 1 mismatch. bottom: prediction of off-targets with 2 mismatches (data in upper triangle/ model in lower triangle). **(C)** Comparison of model to NucleaSeq data. top: prediction of off-targets with 1 mismatch. bottom: prediction of off-targets with 2 mismatches (data in upper triangle/ model in lower triangle). **(D)** Free-energy landscape for 1nM sgCas9-RNA interaction with on-target DNA. Green curves represent fit results from individual runs of our Simulated Annealing optimization algorithm whose resulting values differ less than 5% from the best-solution's outcomes (figure A) (see **S.I.**). Black shows median to guide the eye. Pink shows best solution, used to produce figures A-C. **(E)** Mismatch penalties as a function of position along the RNA-DNA hybrid. Blue dots show individual fit results (after selection). Black shows median to guide the eye. Pink shows best solution, used to produce figures A-C. **(F)** Forward rate parameters. Green dots show individual fit results (after selection). Black shows median to guide the eye. Pink shows best solution, used to produce figures A-C.

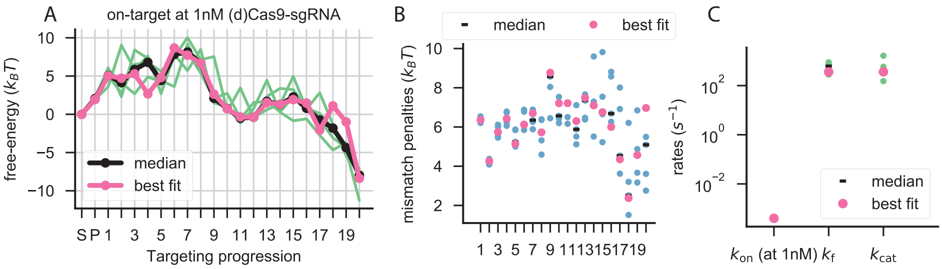


Figure S3.2: related to Figures ?? and ??. Simultaneous fit to HiTS-FLIP and NucleaSeq data. (A) Free-energy landscape for 1nM sgCas9-RNA interaction with on-target DNA. Green curves represent fit results from individual runs of our Simulated Annealing optimization algorithm whose resulting values differ less than 5% from the best-solution's outcomes (**Figures ??**A-B) (see **S.I.**). Black shows median to guide the eye. Pink shows best solution, used to produce **Figures ??**A-C. **(B)** Mismatch penalties as a function of position along the RNA-DNA hybrid. Blue dots show individual fit results (after selection). Black shows median to guide the eye. Pink shows best solution, used to produce **Figures ??**A-C. **(C)** Forward rate parameters. Green dots show individual fit results (after selection). Black shows median to guide the eye. Pink shows best solution, used to produce **??**A-C.

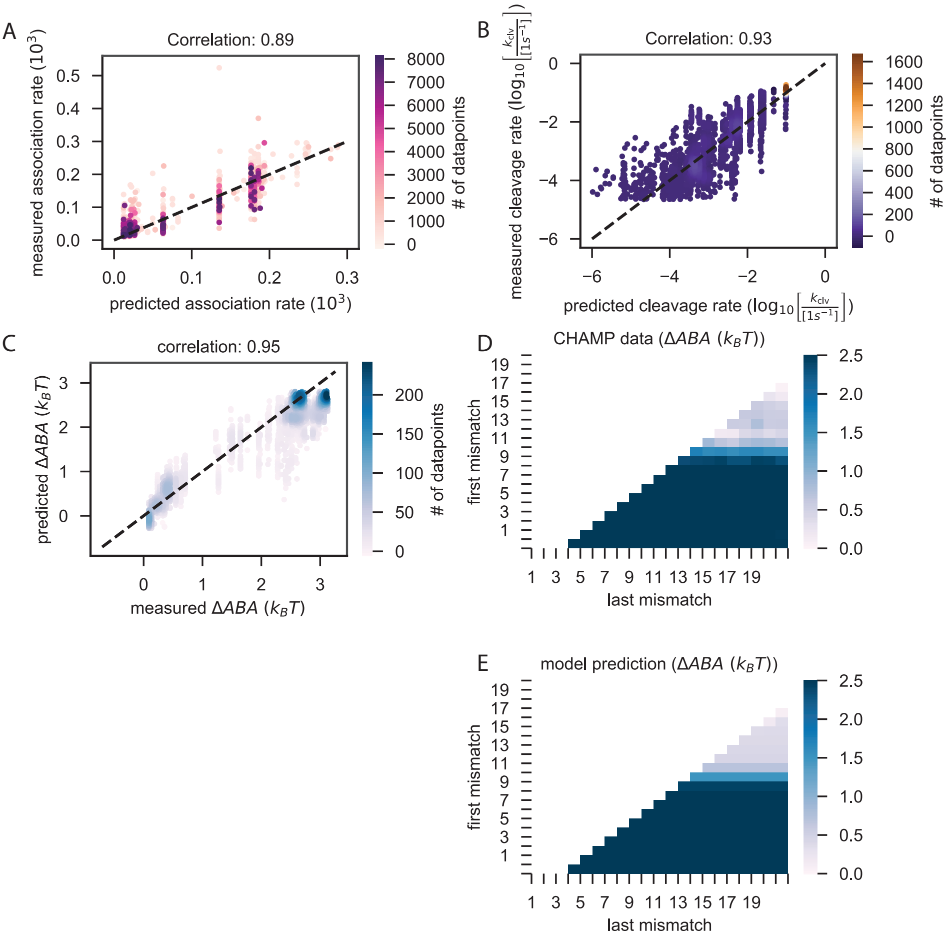


Figure S3.3: related to Figure ??. Correlation plots and additional test data. **(A)** Correlation of model values (fit+prediction) to HiTS-FLIP data. After making a 2D histogram of the data points, each is assigned a color according to the histogram's bin wherein they lie. Darker color indicates a higher density of data points. Dashed line indicates perfect correlation. Both data fitted against (up until 2 mismatches) and the remainder of the library (>2 mismatches) are included. The latter therefore serves as a test. **(B)** Correlation of model values (fit) to NucleaSeq data. Orange/Purple indicates a higher/lower density of data points. **(C)** Correlation of model values (prediction) to CHAMP data. Darker color indicates a higher density of data points. **(D)** CHAMP data for off-targets with consecutive mismatches. Values on the vertical/horizontal axis indicate the first/final mismatch in the stretch. **(E)** Model prediction of the data shown in figure D.

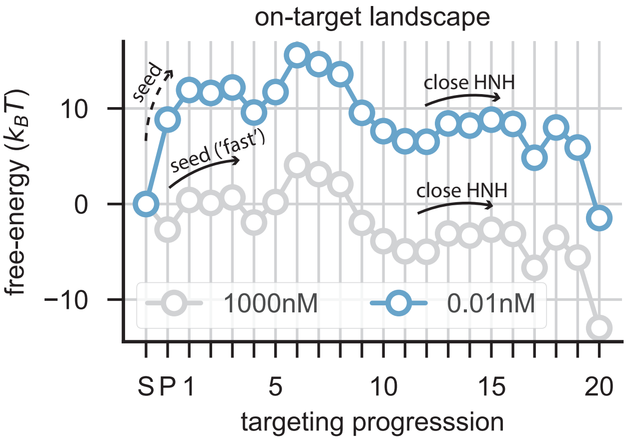


Figure S3.4: related to Figure ??. Free-energy landscapes at varying nuclease concentrations. Free-energy landscape for 0.001nM (blue) and 100nM (grey) Cas9-sgRNA interacting with on-target DNA. The height of the first effective barrier is modulated by nuclease concentration, while the height of the second remains constant. Hence, at higher nuclease concentrations, the difference between dCas9 binding and Cas9 cleavage rates is more pronounced.

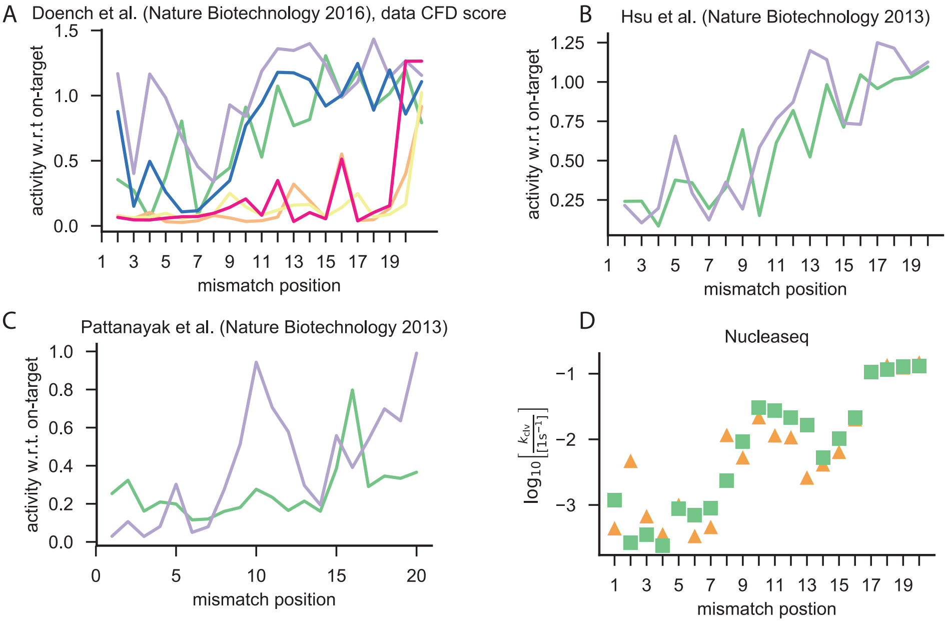


Figure S3.5: related to Figure ??. comparing single-mismatch profiles for various guides (data taken across the literature). (A) Cleavage activity w.r.t on-target, for different guides. Data from Doench et al. [**?**] (processed dataset from Zhang et al. [**?**]). **(B)** Cleavage activity w.r.t on-target, for different guides. Data from Hsu et al. [**?**] (processed dataset from Zhang et al. [**?**]). **(C)** Cleavage activity w.r.t on-target, for different guides. Data from Pattanayak et al. [**?**]. **(D)** NucleaSeq data for guide used throughout this study (orange triangles) and a second guide (green squares).

References

[] J. D. Sander and J. K. Joung, *CRISPR-Cas systems for editing, regulating and targeting genomes,* Nature Biotechnology **32**, 347 (2014).

[] H. Wang, M. La Russa, and L. S. Qi, *CRISPR/Cas9 in Genome Editing and Beyond,* Annual Review of Biochemistry **85**, 227 (2016).

[] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, *A Programmable Dual-RNA – Guided,* Science **337**, 816 (2012), arXiv:38 .

[] L. A. Gilbert, M. H. Larson, L. Morsut, Z. Liu, G. A. Brar, S. E. Torres, N. Stern-Ginossar, O. Brandman, E. H. Whitehead, J. A. Doudna, W. A. Lim, J. S. Weissman, and L. S. Qi, *CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes,* Cell **154**, 442 (2013).

[] B. Chen, L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G. W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, and B. Huang, *Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system,* Cell **155**, 1479 (2013).

[] W. K. Spoelstra, J. M. Jacques, F. L. Nobrega, A. C. Haagsma, M. Dogterom, T. Idema, S. J. Brouns, and L. Reese, *CRISPR-based DNA and RNA detection with liquid phase separation,* Bioarxiv , 1 (2018).

[] X. Wang, E. Xiong, T. Tian, M. Cheng, W. Lin, and J. Sun, *CASLFA : CRISPR / Cas9-mediated lateral flow nucleic acid assay,* Bioarxiv (2019).

[] L. Amoasii, H. Li, E. Sanchez-Ortiz, A. Mireault, D. Caballero, R. Bassel-Duby, E. N. Olson, J. C. Hildyard, R. Harron, C. Massey, R. J. Piercy, T.-R. Stathopoulou, and J. M. Shelton, *Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy,* Science **362**, 86 (2018).

[] C. Y. Park, D. H. Kim, J. S. Son, J. J. Sung, J. Lee, S. Bae, J. H. Kim, D. W. Kim, and J. S. Kim, *Functional Correction of Large Factor VIII Gene Chromosomal Inversions in Hemophilia A Patient-Derived iPSCs Using CRISPR-Cas9,* Cell Stem Cell **17**, 213 (2015).

[] F. Jiang, D. W. Taylor, J. S. Chen, J. E. Kornfeld, K. Zhou, A. J. Thompson, E. Nogales, and J. A. Doudna, *Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage,* Science **351**, 867 (2016), arXiv:arXiv:1011.1669v3 .

[] Y. S. Dagdas, J. S. Chen, S. H. Sternberg, J. A. Doudna, and A. Yildiz, *A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9,* Science Advances **3**, 1 (2017).

[] S. H. Sternberg, B. Lafrance, M. Kaplan, and J. A. Doudna, *Conformational control of DNA target cleavage by CRISPR-Cas9,* Nature **527**, 110 (2015).

[] M. Yang, S. Peng, R. Sun, J. Lin, N. Wang, and C. Chen, *The Conformational Dynamics of Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET,* Cell Reports **22**, 372 (2018).

[] E. A. Boyle, J. O. L. Andreasson, L. M. Chircus, S. H. Sternberg, M. J. Wu, C. K. Guegler, J. A. Doudna, and W. J. Greenleaf, *High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding,* Proceedings of the National Academy of Sciences **114**, 5461 (2017).

[] J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten, and D. E. Root, *Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9,* Nature Biotechnology **34**, 184 (2016), arXiv:15334406 .

[] Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung, and J. D. Sander, *High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells,* Nature Biotechnology **31**, 822 (2013), arXiv:NIHMS150003 .

[] P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, T. J. Cradick, L. A. Marraffini, G. Bao, and F. Zhang, *DNA targeting specificity of RNA-guided Cas9 nucleases,* Nature Biotechnology **31**, 827 (2013), arXiv:NIHMS150003 .

[] S. K. Jones Jr, J. A. Hawkins, N. V. Johnson, C. Jung, K. Hu, R. James, J. S. Chen, J. A. Doudna, W. H. Press, and I. J. Finkelstein, *Massively parallel kinetic profiling of natural and engineered CRISPR nucleases,* BioRxiv , 1 (2019).

[] D. Kim, K. Luk, S. A. Wolfe, and J.-S. Kim, *Evaluating and Enhancing Target Specificity of Gene-Editing Nucleases and Deaminases,* Annual Review of Biochemistry , 1 (2019).

[] V. Pattanayak, S. Lin, J. P. Guilinger, E. Ma, J. A. Doudna, and D. R. Liu, *High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity,* Nature Biotechnology **31**, 839 (2013).

[] S. Q. Tsai and J. K. Joung, *Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases,* Nature Reviews Genetics **17**, 300 (2016).

[] G. Cullot, J. Boutin, J. Toutain, F. Prat, P. Pennamen, C. Rooryck, M. Teichmann, E. Rousseau, I. Lamrissi-Garcia, V. Guyonnet-Duperat, A. Bibeyran, M. Lalanne, V. Prouzet-Mauléon, B. Turcq, C. Ged, J. M. Blouin, E. Richard, S. Dabernat, F. Moreau-Gaudry, and A. Bedel, *CRISPR-Cas9 genome editing induces* *megabase-scale chromosomal truncations,* Nature Communications **10**, 1 (2019).

[] C. Kuscu, S. Arslan, R. Singh, J. Thorpe, and M. Adli, *Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease,* Nature Biotechnology **32**, 677 (2014), arXiv:arXiv:1208.5721 .

[] H. O'Geen, I. M. Henry, M. S. Bhakta, J. F. Meckler, and D. J. Segal, *A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture,* Nucleic Acids Research **43**, 3389 (2015).

[] X. Wu, D. A. Scott, A. J. Kriz, A. C. Chiu, P. D. Hsu, D. B. Dadon, A. W. Cheng, A. E. Trevino, S. Konermann, S. Chen, R. Jaenisch, F. Zhang, and P. A. Sharp, *Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells,* Nature Biotechnology **32**, 670 (2014), arXiv:NIHMS150003 .

[] M. Stemmer, T. Thumberger, M. Del Sol Keyer, J. Wittbrodt, and J. L. Mateo, *CCTop: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool,* PLoS ONE **10**, 1 (2015).

[] G. Chuai, H. Ma, J. Yan, M. Chen, N. Hong, D. Xue, C. Zhou, C. Zhu, K. Chen, B. Duan, F. Gu, S. Qu, D. Huang, J. Wei, and Q. Liu, *DeepCRISPR: Optimized CRISPR guide RNA design by deep learning,* Genome Biology **19**, 1 (2018).

[] J. Listgarten, M. Weinstein, B. P. Kleinstiver, A. A. Sousa, J. K. Joung, J. Crawford, K. Gao, L. Hoang, M. Elibol, J. G. Doench, and N. Fusi, *Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs,* Nature Biomedical Engineering **2**, 38 (2018).

[] I. Farasat and H. M. Salis, *A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation,* PLoS Computational Biology **12**, 1 (2016).

[] D. Zhang, T. Hurst, D. Duan, and S.-J. Chen, *Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design,* Proceedings of the National Academy of Sciences **116**, 8693 (2019).

[] M. Klein, B. Eslami-Mossallam, D. Arroyo, and M. Depken, *Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules,* Cell Reports **22** (2018), 10.1016/j.celrep.2018.01.045.

[] C. Anders, O. Niewoehner, A. Duerst, and M. Jinek, *Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease,* Nature **513**, 569 (2014), arXiv:NIHMS150003 .

[] F. Jiang, K. Zhou, S. Gressel, and J. A. Doudna, *A cas9 guide RNA complex preorganized for target DNA recognition,* Science **348**, 1477 (2015).

[] E. A. Josephs, D. D. Kocak, C. J. Fitzgibbon, J. McMenemy, C. A. Gersbach, and P. E. Marszalek, *Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage,* Nucleic Acids Research **43**, 8924 (2015).

[] M. Rutkauskas, T. Sinkunas, I. Songailiene, M. S. Tikhomirova, V. Siksnys, and R. Seidel, *Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection,* Cell Reports **10**, 1534 (2015).

[] M. D. Szczelkun, M. S. Tikhomirova, T. Sinkunas, G. Gasiunas, T. Karvelis, P. Pschera, V. Siksnys, and R. Seidel, *Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes,* Proceedings of the National Academy of Sciences **111**, 9798 (2014).

[] Y. Xiao, M. Luo, R. P. Hayes, J. Kim, S. Ng, F. Ding, M. Liao, and A. Ke, *Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System,* Cell **170**, 48 (2017).

[] N. Bisaria, I. Jarmoskaite, and D. Herschlag, *Lessons from Enzyme Kinetics Reveal Specificity Principles for RNA-Guided Nucleases in RNA Interference and CRISPR-Based Genome Editing,* Cell Systems **4**, 21 (2017).

[] K. Sung, J. Park, Y. Kim, N. K. Lee, and S. K. Kim, *Target Specificity of Cas9 Nuclease via DNA Rearrangement Regulated by the REC2 Domain,* Journal of the American Chemical Society **140**, 7778 (2018).

[] C. Jung, J. A. Hawkins, S. K. Jones, Y. Xiao, J. R. Rybarski, K. E. Dillard, J. Hussmann, F. A. Saifuddin, C. A. Savran, A. D. Ellington, A. Ke, W. H. Press, and I. J. Finkelstein, *Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips,* Cell **170**, 35 (2017).

[] T. Künne, D. C. Swarts, and S. J. Brouns, *Planting the seed: Target recognition of short guide RNAs,* Trends in Microbiology **22**, 74 (2014).

[] E. Semenova, M. M. Jore, K. A. Datsenko, A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. J. Brouns, and K. Severinov, *Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence,* Proceedings of the National Academy of Sciences **108**, 10098 (2011).

[] Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio, and J. K. Joung, *Improving CRISPR-Cas nuclease specificity using truncated guide RNAs,* Nature Biotechnology **32**, 279 (2014), arXiv:29 .

[] M. Jinek, F. Jiang, D. W. Taylor, S. H. Sternberg, E. Kaya, E. Ma, C. Anders, M. Hauer, K. Zhou, S. Lin, M. Kaplan, A. T. Iavarone, E. Charpentier, E. Nogales, and J. A. Doudna, *Structures of Cas9 endonucleases reveal RNA-mediated conformational activation,* Science **343** (2014), 10.1126/science.1247997.

[] J. S. Chen, Y. S. Dagdas, B. P. Kleinstiver, M. M.

Welch, A. A. Sousa, L. B. Harrington, S. H. Sternberg, J. K. Joung, A. Yildiz, and J. A. Doudna, *Enhanced proofreading governs CRISPR-Cas9 targeting accuracy,* Nature **550**, 407 (2017).

[] M. Haeussler, K. Schönig, H. Eckert, A. Eschstruth, J. Mianné, J. B. Renaud, S. Schneider-Maunoury, A. Shkumatava, L. Teboul, J. Kent, J. S. Joly, and J. P. Concordet, *Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR,* Genome Biology **17**, 1 (2016).

[] J. S. Chen, E. Ma, L. B. Harrington, M. Da Costa, X. Tian, J. M. Palefsky, and J. A. Doudna, *CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity,* Science **360**, 436 (2018).

[] J. S. Gootenberg, O. O. Abudayyeh, J. W. Lee, P. Essletzbichler, A. J. Dy, J. Joung, V. Verdine, N. Donghia, N. M. Daringer, C. A. Freije, C. Myhrvold, R. P. Bhattacharyya, J. Livny, A. Regev, E. V. Koonin, D. T. Hung, P. C. Sabeti, J. J. Collins, and F. Zhang, *Nucleic acid detection with CRISPR-Cas13a/C2c2.* Science (New York, N.Y.) **356**, 438 (2017), arXiv:15334406 .

[] J. S. Gootenberg, O. O. Abudayyeh, M. J. Kellner, J. Joung, J. J. Collins, and F. Zhang, *Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6.* Science **444**, 439 (2018).

[] L. B. Harrington, J. S. Chen, E. Ma, I. P. Witte, J. C. Cofsky, J. A. Doudna, D. Burstein, J. F. Banfield, D. Paez-Espino, and N. C. Kyrpides, *Programmed DNA destruction by miniature CRISPR-Cas14 enzymes,* Science **362**, 839 (2018).

[] D. Kim, J. Kim, J. K. Hur, K. W. Been, S. H. Yoon, and J. S. Kim, *Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells,* Nature Biotechnology **34**, 863 (2016).

[] B. P. Kleinstiver, S. Q. Tsai, M. S. Prew, N. T. Nguyen, M. M. Welch, J. M. Lopez, Z. R. McCaw, M. J. Aryee, and J. K. Joung, *Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells,* Nature Biotechnology **34**, 869 (2016), arXiv:15334406 .

[] B. P. Kleinstiver, V. Pattanayak, M. S. Prew, S. Q. Tsai, N. T. Nguyen, Z. Zheng, and J. K. Joung, *High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects,* Nature **529**, 490 (2016), arXiv:9605103 [cs] .

[] I. M. Slaymaker, L. Gao, B. Zetsche, D. A. Scott, W. X. Yan, and F. Zhang, *Rationally engineered Cas9 nucleases with improved specificity,* Science **351**, 84 (2016), arXiv:NIHMS150003 .

[] N. Amrani, X. D. Gao, P. Liu, A. Edraki, A. Mir, R. Ibraheim, A. Gupta, K. E. Sasaki, T. Wu, P. D. Donohoue, A. H. Settle, A. M. Lied, K. McGovern, C. K. Fuller, P. Cameron, T. G. Fazzio, L. J. Zhu, S. A. Wolfe, and E. J. Sontheimer, *NmeCas9 is an intrinsically high-fidelity genome-editing platform Jin-Soo Kim,* Genome Biology **19**, 1 (2018).

[] W. R. Becker, B. Ober-Reynolds, K. Jouravleva, S. M. Jolly, P. D. Zamore, and W. J. Greenleaf, *High-Throughput Analysis Reveals Rules for Target RNA Binding and Cleavage by AGO2,* Molecular Cell , 1 (2019).

[] A. Tambe, A. East-Seletsky, G. J. Knott, J. A. Doudna, and M. R. O'Connell, *RNA Binding and HEPN-Nuclease Activation Are Decoupled in CRISPR-Cas13a,* Cell Reports **24**, 1025 (2018).

[] F. Heigwer, G. Kerr, and M. Boutros, *E-CRISP: Fast CRISPR target site identification,* Nature Methods **11**, 122 (2014).

[] K. Labun, T. G. Montague, J. A. Gagnon, S. B. Thyme, and E. Valen, *CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering,* Nucleic acids research **44**, W272 (2016).

[] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi Jr., *Optimization by simulated annealing,* Science **220**, 671 (1983).

II

Target search

4

Why Argonaute is needed to make microRNA target search fast and reliable

MicroRNA (miRNA) interferes with the translation of cognate messenger RNA (mRNA) by finding, preferentially binding, and marking it for degradation. To facilitate the search process, Argonaute (Ago) proteins come together with miRNA, forming a dynamic search complex. In this review we use the language of free-energy landscapes to discuss recent single-molecule and high-resolution structural data in the light of theoretical work appropriated from the study of transcription-factor search. We suggest that experimentally observed internal states of the Ago-miRNA search complex may have the explicit biological function of speeding up search while maintaining specificity.

This chapter has been published as: **M.Klein***, S.D.Chandradoss*, M.Depken and C.Joo. Why Argonaute is needed to make microRNA target search fast and reliable. *Seminars in Cell and Developmental Biology* **65** 20–28 (2017) (*co-first authors)

4.1. Introduction

Eukaryotes regulate gene expression post-transcriptionally through the RNA interference (RNAi) pathway. This pathway begins with the transcription of non-coding RNA and its subsequent maturation into microRNA (miRNA). To facilitate search and suppression of target messenger RNA (mRNA), Argonaute (Ago) proteins join together with the miRNA molecule, forming an efficient search complex [**?** **?**]. In the pool of cellular RNA, the search complex finds mRNA cognate to its miRNA and primes its degradation. As the search relies on thermal motion, the functioning of the search complex can be understood in terms of diffusion and the binding-energy landscape of mRNA-Ago-miRNA interactions. In this Review, we discuss recent single-molecule and structural data on Ago, and borrow free-energy considerations and theory from transcription-factor search, highlighting how several of the observed Ago conformations could function to speed up the search process.

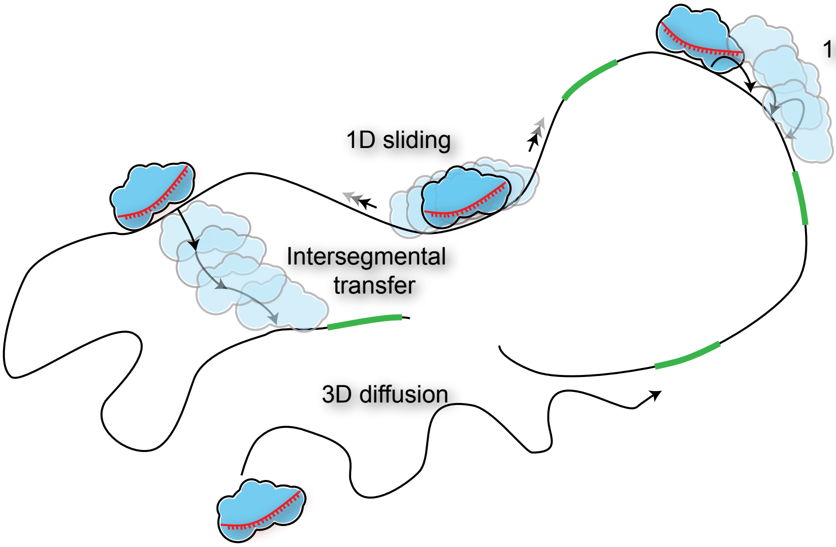


Figure 4.1: Facilitated diffusion. Four different modes of search can in principle be distinguished. 1) 3D search: An Argonaute protein probes a new sequence by first unbinding, then diffusing through the cytosol, and finally binding to probe a new uncorrelated site. 2) Sliding: A non- specifically bound protein laterally diffuses along the mRNA to probe a new site, probing every potential intermediate site from the start to the new site. 3) Hopping: A non-specifically bound protein unbinds, but quickly rebinds again to a site close by (along the RNA) from where it unbound, but not necessarily probing every site in between. 4) Intersegmental transfer: a hopping mechanism where unbinding and binding positions are correlated in 3D space, but far apart along the RNA. This is possible due to the coiled conformation RNA adapts *in vivo.* binding

4.2. Target search in 1D and 3D

Ever since the initial observations of an astonishingly high association rate of the E. coli Lac repressor to the lac operon [**?**], researchers have been trying to understand general mechanisms that could speed up target search on nucleic-acid templates. In their seminal work [**?**], Berg, Winter and von Hippel proposed a facilitated diffusion mechanism by which the protein combines three-dimensional diffusion through the cytoplasm with lateral diffusion along the DNA (see Fig.**??**) [**?**]. We here qualitatively summarize the theoretical arguments behind this suggestion and review the experimental evidence for lateral diffusion by various search complexes.

4.2.1. Facilitated diffusion enables rapid target search of miRNA

Though facilitated diffusion was originally aimed at transcription-factor search on DNA, the same arguments apply to any searcher along a nucleic acid sequence, including Ago-miRNA search on RNA. The benefit of employing both 3D and 1D search can be qualitatively understood as follows: To find the next sequence to probe, it will always be faster to diffuse a short distance laterally along the RNA (through hopping and sliding; Fig. **??**) than to diffuse a long distance through the cytosol. As lateral diffusion brings you to close-by sites, there exists a point beyond which the search complex starts predominantly probing sites already visited. At this point it becomes favorable to move to an unprobed RNA neighborhood by diffusing through the cytosol. Minimizing redundancy of the one-dimensional (1D) search thus comes at the cost of employing the slower 3D search, and there exists an optimum partitioning between the two [**? ? ? ? ?**].

4.2.2. Experimental evidence for lateral diffusion during target search

Single-molecule fluorescence studies brought direct evidence of lateral diffusion during molecular target search, including sliding of transcription factors [**? ?**], DNA repair proteins [**? ? ?**] zinc-finger proteins [**?**], and the DNA recombination protein RecA [**?**]. Like Argonaute, RecA makes a nucleoprotein complex (a RecA—single-stranded DNA filament) that is ready to basepair for target search [**? ? ? ? ?**]. In order to investigate lateral diffusion of Ago-miRNA on RNA, we adopted an *in vitro* single-molecule FRET assay that was developed for studying RecA-mediated target search [**?**]. We placed two identical binding sites on a single target RNA strand, each of which led to a different FRET efficiency with Ago-miRNA bound [**?**]. We observed that a substantial fraction of the binding events ($> 50\%$) shuttled between two strong binding positions via rapid lateral diffusion. When using a volume-occupying reagent (PEG) to mimic physiological conditions, most binding events ($> 90\%$) displayed shuttling by the same Ago-miRNA complex. This suggests that lateral diffusion could also be important for *in vivo* microRNA search.

4.3. Multiple protein configurations for fast lateral diffusion and stable target recognition

While target search is sped up by facilitated diffusion, Slutsky and Mirny [**? ?**] argued that it is not possible to have both fast lateral diffusion and stable/preferential binding to the target using a single nucleoprotein conformation. The more stable binding to the target is, the more stable binding to similar sequences also becomes, and the lateral diffusion slows down as it gets increasingly trapped at non-target sites. To understand what is needed for the resolution of this apparent paradox, we now follow Slutsky and Mirny [**? ?**] and consider the statistical variation of binding energies along the substrate (which for us is mRNA).

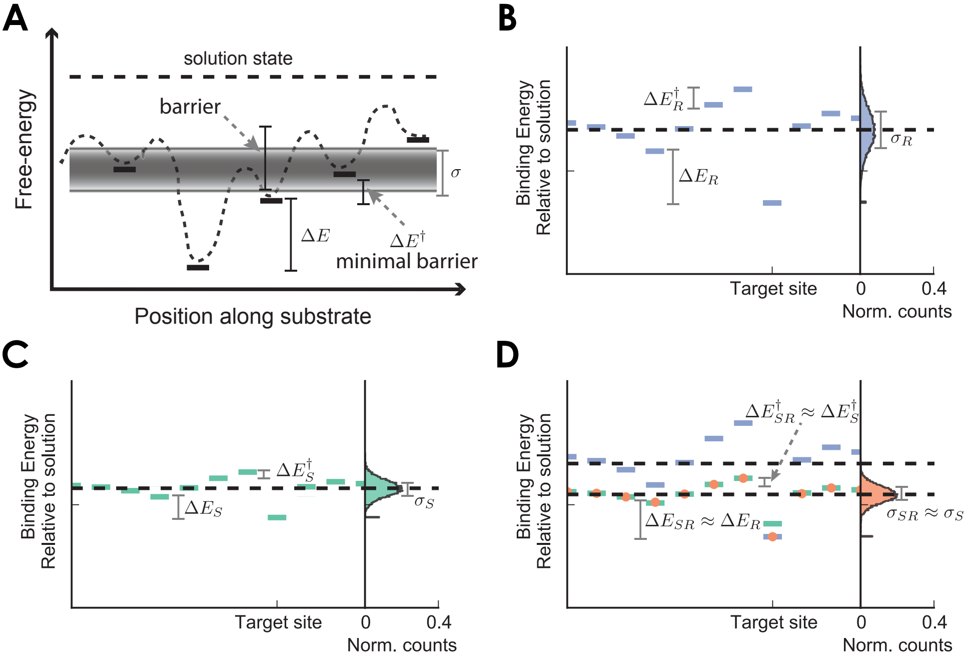


Figure 4.2: Search-stability paradox. **(A)** Energies of the binding sites are shown as short black horizontal markers. Being a sum of base pairing energies, binding energies are (approximately) Gaussian distributed with a standard deviation σ. The target site is separated from the other binding sites by an energy of about ΔE. When diffusing laterally, the minimal barrier towards diffusion is set by the energetic difference between neighbouring sites ($\Delta E^{\dagger}$). In reality there are intervening barriers, as depicted by the dashed line. With little loss of generality, we will ignore these additional contributions to the barriers and focus on the best-case scenario. **(B)** Recognition mode – Stable binding, but slow search: A larger difference between target and non-target energies comes at the cost of having larger barriers towards diffusion. The right panel shows the complete distribution of energetic states (standard deviation σ_{R}) of which a subset is plotted in the left panel. The typical (minimal) barrier towards diffusion ($\Delta E_{\mathrm{R}}^{\dagger}$) and differential binding energy (ΔE_{R}) are indicated. **(C)** Search mode – Fast search, but no stable binding: Decreasing the barriers also decreases the difference between target and non-target energy, which hampers the ability of the search complex to selectively bind to the target. The right panel shows the complete distribution of energetic states (standard deviation σ_{S}) of which a subset is plotted in the left panel. The typical (minimal) barrier towards diffusion ($\Delta E_{\mathrm{S}}^{\dagger}$) and differential binding energy (ΔE_{S}) are indicated. **(D)** Search + Recognition - Fast search and stable binding: If the search complex posesses (at least) two distinct binding modes, it becomes possible to combine the landscapes of figures B (blue) and C (green) to enable rapid diffusion ($\Delta E^{\dagger} \approx \Delta E_{\mathrm{S}}^{\dagger}$) towards the target without loss of selectivity ($\Delta E \approx \Delta E_{\mathrm{R}}$) (orange).

4.3.1. Resolving the speed-stability paradox by utilizing multiple binding modes

Apart from the target, the sequences being searched through can be considered as essentially random and uncorrelated [**? ?**]. A substantially preferential binding to the target requires that a correct match has a considerable energetic difference (ΔE, for definition see Fig. **??**A) to all partial matches. Slutsky and Mirny assume that the search complex has a binding energy roughly proportional to the degree of sequence homology between probed and target sequence. Under the assumption that the binding energy comes only from individual nucleotide-basepairing energies, a large energetic difference between target and non-target positions can only be achieved by large differences in pairing for each nucleotide. A general increase of basepairing energies results in a larger standard deviation among binding energies at different positions (compare σ_{R} of the “recognition” landscape and σ_{S} of the “search” landscape in Fig. **??**B and C respectively), and the diffusion constant along the mRNA can be shown to decrease sharply [**? ?**]. In Fig. **??**B we illustrate how a large recognition energy will generally imply large barriers to lateral diffusion ($\Delta E^{\dagger}$, for definition see Fig. **??**A),resulting in a slow search process. Reversely, in Fig. **??**C we illustrate how small barriers to diffusion implies poor recognition. Slutsky and Mirny proposed that the coupling between recognition energy and diffusion barrier ($\Delta E^{\dagger}$ being proportional to ΔE) can be broken if the search complex can stochastically switch between two internal modes with different binding energy strength (Fig. **??**D):

1. A search (S) mode: small affinity differences and fast diffusion ($\sigma_{\mathrm{S}} \lessapprox 2k_BT$; Ref. [**?**])

2. A recognition (R) mode: large affinity differences and slow diffusion ($\sigma_{\mathrm{R}} \gtrapprox 5k_BT$; Ref. [**?**])

An efficient searcher must have evolved the ability to combine the search and recognition modes. Thereby, the non-specific (average) energies (dashed lines in Fig. **??**B-D) are arranged such that all energies of the search mode lie between the energies of all non-target sites and the target in the recognition mode (see Fig. **??**D). Such systems predominantly move according to the search mode when not at the target site, but predominantly occupy the recognition mode once at the target (see states with orange dots in Fig. **??**D). The effective search barriers are now set by the search mode ($\Delta E^{\dagger} \approx \Delta E^{\dagger}_{\mathrm{S}}$) while the recognition energies are set by the recognition mode ($\Delta E \approx \Delta E_{\mathrm{R}}$). Both fast search and stable recognition is thus in principle possible if the searching protein possesses at least two distinct binding modes, and the above case represents the theoretical ideal scenario (for more general cases see [**? ? ? ? ?**]).

4.3.2. Experimental evidence for two initial binding modes of Ago-miRNA

Both recent structural and single-molecule data of eukaryotic Ago proteins suggest that the hybridization between guide and target is gradual and is coupled to structural changes in the search complex. We here discuss these studies in the light of a search-stability paradox for Ago-miRNA.

Biochemical, structural and computational analyses suggest that Argonaute divides its miRNA guides into five functional domains (5’anchor, seed, mid region, 3’ supplementary region, and the tail region) (Fig. **??**). The seed region (nt 2–8) is crucial for gene suppression [**? ? ? ? ? ? ?**], and it was shown that protein mediated interactions stabilize nt 2–6 into an A-form-helix that exposes nt 2–4 (or 2–5) for base paring with the target (Fig. **??**A) [**?**]. Based on this observation, Schirle et al [**?**] proposed a step-wise target recognition for human Argonaute-2 (hAgo2), in which the initial recognition of the target occurs in the 5’ part of the miRNA. Two subsequent single-molecule studies showed that Ago-miRNA indeed uses this so-called sub-seed for the initial weak recognition. Solomon et al designed

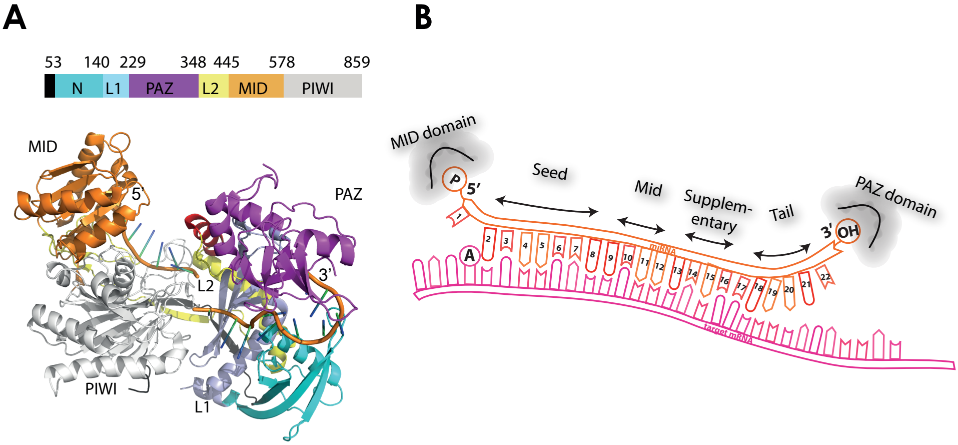


Figure 4.3: Structural and domain overview of hAgo2 and miRNA. (A) The binary structure of hAgo2-miRNA showing four well conserved domains among Argonaute proteins (snapshot of the structure 4W5N taken in pymol) **(B)** Argonaute proteins divide miRNA(orange) in to several domains. The 5' phosphate and nt 1 of miRNA (anchor) is bound to the pocket in the MID domain. The nt 2–8 are known as seed sequence, as they are crucial for initial targeting. The nt 9–10 have the least significant role in target recognition and are known as the mid region. The 3' supplementary region is comprised of nt 13–16, they also have considerable role in stabilizing miRNA-target interaction. The nucleotides beyond the 16th do not base pair with the target and are called the tail region. The 3' OH is bound to the binding pocket in PAZ domain making it as a 3' anchor. The t1 Adenosine (t1A) in the target RNA (pink) binds to the binding pocket in MID domain..

di-nucleotide mutation constructs for mouse Ago-miRNA and measured the unbinding rate from the target RNA [**?**]. We have also shown that, when the paired region was gradually shrunk from the full seed (nt 2-8) to only the first three nucleotides (nt 2-4), no difference in the binding rate was noticeable [**?**]. These two results showed that it is only the first three nucleotides of the seed that are used to maintain weak interaction during the initial search.

The two single-molecule works also suggested that Ago-miRNA exhibits a sharp increase in the binding affinity when the number of paired nucleotides changes from 6 to 7 [**? ?**]. Comparison of crystal structures suggests that this property originates from the fact that Argonaute makes the guide kink away from the A-form stacked structure in several places [**? ? ? ?**]. The most prominent kink disrupting the helical arrangement of the guide is between nt 6 and 7 (Fig. **??**B). Base paring to the target, therefore, requires a shift of the helix-7 that clashes with the incoming target. After pairing of nt 2-4, hAgo2 undergoes a conformational change leading to a 4Å displacement of the helix-7 loop and allowing base pairing of nt 6–8 (Fig. **??**C). It was hypothesized that the sharp increase in the time bound between having 6 and 7 nt matching is caused by the conformational change of the helix-7 motif [**?**]. We here suggest that Ago makes a change from a weak binding (search) mode using nt 2- 4 to a strong binding (recognition) mode using a full seed through the conformational change of the helix-7.

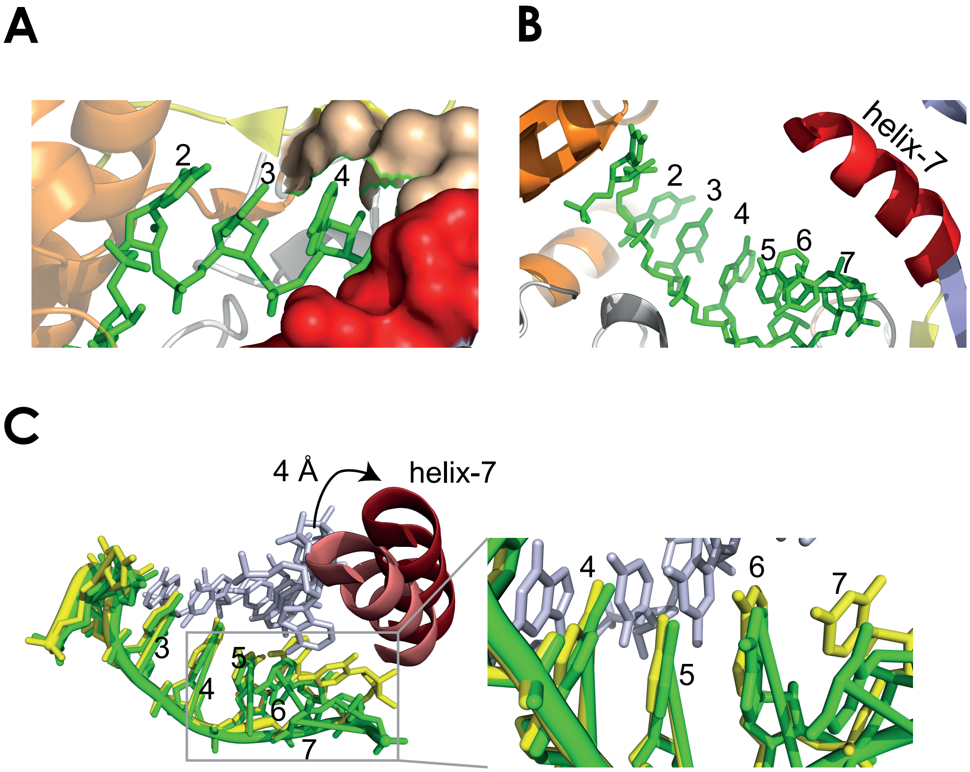


Figure 4.4: Seed of miRNA and hAgo2-helix7. **(A)** Nucleotides 2–4 (green) of the guide RNA are well exposed by residues in the PIWI domain (golden surface) possibly for initial target recognition (snapshot of the structure 4W5N taken in pymol). (B) **(B)** The access to nt 5–7 of the guide (green) is blocked by the helix-7 motif (red). The base paring of target to guide nt 5–7 would require displacement of helix-7 (snapshot of the structure 4W5N taken in pymol). **(C)** Upon base paring with the target (grey) the helix-7 motif is displaced by 4 Å compared to guide-only structure. The displacement of helix-7 removes the constraints from nt 6 and 7 (yellow) compared to guide only structure (green) making nt 6 and 7 available for base paring (see the close-up view in the right panel). (snapshot of the structures 4W5N (guide only) and 4W5O (guide and target) taken in pymol).

4.3.3. The experimental evidence for additional binding modes of Ago-miRNA

In addition to the helix-7 movement, more conformational changes take place after seed pairing is achieved, and before the bound Ago-miRNA complex becomes cleavage competent. First, binding of the supplementary region (nt 13-16) ensuing the seed pairing enhances the binding stability of Ago-miRNA [**?**]. But the pairing beyond nt 8 is restricted by a physical constraint [**?**](Fig. **??**A). Widening up of a channel between PAZ and N-terminus domains allows for a rearrangement of the disordered supplementary region (nt 13-16) of the miRNA into a helical A-form, preparing it for pairing with the target RNA (Fig. **??**B)[**?**]. It remains to be seen whether target recognition is enhanced by this additional checkpoint. Second, biochemical and single-molecule studies have shown that the base paring in the mid region is necessary for cleavage of target RNA [**? ?**]. But Jo et al also observed that a

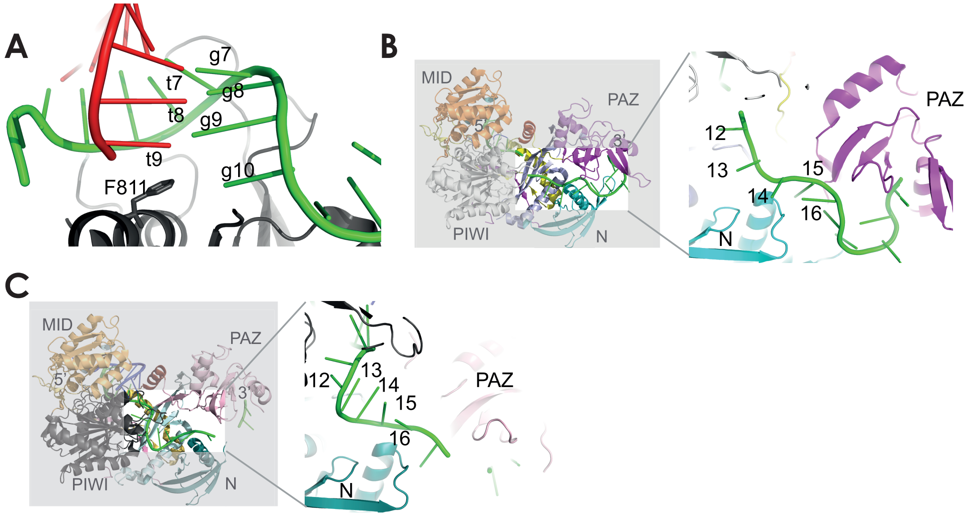


Figure 4.5: Cleavage competent state. **(A)** Structure showing the base pairing between a guide strand (green) and a target strand (red). The base pairing beyond nt 8(g8) is blocked by a residue F811 in a helix of the PIWI domain (snapshot of the structure 4W5O taken in pymol). **(B)** A binary structure of hAgo2-miRNA showing the disordered 3' supplementary region of guide RNA (green) passing through a channel between N domain (blue) and PAZ domain (purple) (snapshot of the structure 4W5N taken in pymol). **(C)** A ternary structure of hAgo2-miRNA and its target showing an A-form helical arrangement of the 3' supplementary region of guide (green) in ternary structure (snapshot of the structure 4W5O taken in pymol). sites

significant portion of Ago-miRNAs were not able to cleave the target RNAs in spite of their perfect complementarity [**?** **?**]. The unsuccessful cleavage of perfect complementary target might be the resultant of a failure to induce an additional conformational change needed for cleavage that involves positioning of Ago's catalytic residues residing near nt 9-10 of the miRNA.

Third, Ago uses its PAZ domain to preclude miRNA from being tightly associated with target RNA. An earlier biochemical study reported that bare RNA as short as 12bp is long enough for stable hybridization (a year of life time) [**?**]. But it was observed that Ago-miRNA (or Ago-guide DNA) often dissociated from its target within seconds to minutes after binding [**?**]. This reversible binding, which is speculated to reduce off-targeting [**?**], is possible because the 3' end of guide RNA is anchored to the PAZ domain and this lowers the binding affinity of Ago-miRNA (especially at the 3' end) to target RNA [**? ? ? ? ? ?**].

In addition to the complex interactions between Ago and a guide strand, a direct interaction between Ago and target RNA also contributes to the target selection. Schirle et al [**?**] showed that hAgo2 interacts with the adenine nucleotide of the target when it is opposite to the 1st nucleotide of the guide. Through a water network, the residues in the MID domain (Fig. **??**A) specifically recognize the t1A anchoring the Ago-miRNA complex to the target. Using a single-molecule assay they showed that t1A does not influences initial target recognition but increases the residence time of Ago-miRNA on to the target RNA, which might enhance its cleavage efficiency [**?**].

4.4. Energy landscape of miRNA target search

Having discussed the evidence that a series of conformational changes are needed to initiate stable binding and cleavage of target mRNA, we now discuss how conformational changes effect the binding-energy landscape. When Ago initially scans the target RNA it exposes only nucleotides 2-4 of the miRNA, termed the sub-seed. In this search mode it does not discriminate strongly based on RNA sequence, and lateral diffusion is likely rapid. A complete match of the sub-seed stabilizes a conformational change that exposes the remainder of the seed (nt 2-8) for base pairing, and, once paired, it slows down the diffusion in this recognition mode (Fig. **??**A). Upon encountering a sequence bearing complementarity to the entire seed, the helix-7 is displaced to allow miRNA to fully pair with the target, and the Ago-miRNA complex arrives in this more stable recognition state (Fig. **??**A and B). We suggest that the function of these various states is analogous to the function of internal states in transcription-factor search (Fig. **??**D).
In figure **??**B we sketch a free-energy landscape of the dominant configuration at varying degrees of base pairing for a perfect match. Transitions requiring conformational changes cost energy, increasing barriers to further base pairing. We construct a sketch of the landscape based on a single-molecule study that reported the existence of various pathways even when the full sequence of miRNA matches with a target [**?**]: a significant fraction of the population showed transient binding (10%) and stable binding with no cleavage (30%). Assuming that the largest barrier to further basepairing originates from the required movement of helix-7, the substantial fraction of transiently binding proteins indicates that this barrier must come close to the barrier to unbind. Further, the even larger fraction of stable but non-cleaving complexes indicates that the average binding energy past helix-7 is strong, and that the cleavage rate is slow compared to experimental times, but fast compared to unbinding.

With these general considerations, we conclude that the free-energy landscape of Fig. **??**B captures at least one search mode (pre-seed pairing) and at least one recognition mode (post-seed pairing). These two modes could be further split up, e.g. the seed pairing into sub-seed and full seed pairing. Still, the general principle behind resolving the speed-stability paradox should apply. To determine the quantitative effects of this energy landscape will require additional theoretical work accounting for gradual base pairing and a series of conformational changes. Using single-molecule techniques and high resolution structural studies, it will also be possible to test the effect of Ago's conformational changes on target search by analysing mutated proteins or directly observe conformational switching (for instance by using FRET such as done for Cas9 in [**?**]).

4.5. Outlook

We have reviewed the principles behind facilitated diffusion and the speed-stability paradox in general target search processes, as well as the experimental evidence for facilitated diffusion in miRNA target search. We further discussed the evidence for multiple search states in the Ago-miRNA search complex, which could help resolve the speed-stability paradox—simultaneously enabling the search to be fast and the binding to the target to be strong.

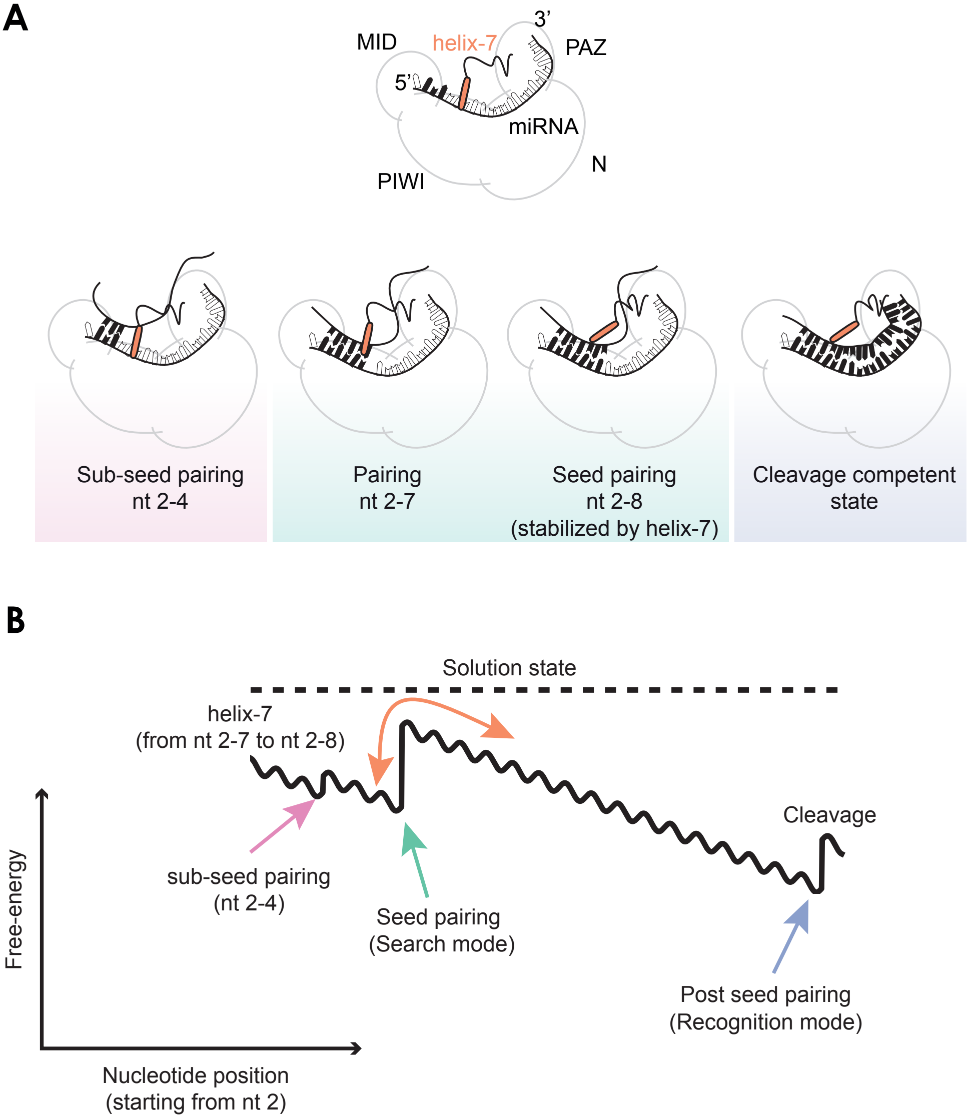


Figure 4.6: Target search process by hAgo2. **(A)**A model summarizing conformational changes during target search by hAgo2-miRNA. In light of the search-stability paradox discussed in Fig. **??** we identify a two search modes (pink + green) and a recognition mode (blue). Alternating between search and recognition modes is enabled through the movement of the helix-7 motif (orange). **(B)** Schematic free-energy diagram for Ago-microRNA target recognition. Forming bonds between target and guide (horizontal axis) makes the complex more stable (vertical axis). In light of the search-stability paradox, as proposed by Slutsky and Mirny and discussed in Fig. **??**, we identify at least 1 search mode (pre-seed pairing, green arrow) and at least one recognition mode (post-seed pairing, blue arrow). To resolve the paradox, Argonaute can use the movement of its helix-7 motif to switch between search and recognition modes (orange arrow). Potentially, additional modes can be distinguished, such as sub-seed pairing (pink arrow).

4.5.1. Further insight into Ago-miRNA target search can improve microRNA target prediction algorithms

Due to the complex nature of the mRNA targeting process, it is far from straightforward to predict what genes are silenced by a particular miRNA. Experimentally, mRNA targets have been found by analysing the effect of miRNA expression on protein production or by performing binding assays [**?**]. For such approaches to work, one needs to know what target gene should be considered from the outset. Using bioinformatics algorithms, potential target sites are scored, and high scoring targets are subsequently tested in experiment. Simple sequence homology between the mRNA to the guiding miRNA does not by itself give an accurate prediction of targets. Presently, typical prediction algorithms are largely phenomenological in nature, for example, assigning higher scores to sequences that fully match the seed of the miRNA and/or are evolutionary conserved. Additionally, accounting for the secondary structure of mRNA and the sequence outside of the targeted 3'-UTR further improves predictions [**? ?**]. A recent combined bioinformatics and *in vivo* study showed that there are at least 14 additional sequence features (for example the length 3'-UTR region and the predicted structural accessibility of the RNA) of the mRNA that improve microRNA target prediction algorithms [**?**]. Yet, despite much effort, prediction algorithms often point to many target sites that cannot be validated experimentally or fail to pick out targets that have been previously validated. Single-molecule studies allow one to study how Ago-miRNA's interaction with RNA binding proteins effects target affinity. Synthesising such molecular level understanding into the free-energy landscapes that we have discussed in this review should help improving the scoring functions of target prediction algorithms by taking the non-equilibrium features of the system into account. Additionally, prediction algorithms can potentially be improved by taking sequences neighbouring the target into account [**? ? ? ?**]. Chandradoss *et al.* showed that, when two identical targets are neighbouring each other, the total retention time was substantially larger than what can be expected on theoretical grounds for two non-interacting targets [**?**]. This synergistic effect might also be observed when a target is neighboured by sub-seed sequences. It will be interesting to determine whether this putative effect exists *in vivo.* Possibly, modelling the physical interaction with neighbouring sites, and accordingly assigning higher scores to those mRNA sequences with a high-density of sub-seed sequences, could then improve target prediction algorithms.

4.5.2. Implications for other target search systems

In the cell, multiple nucleic acid-mediated target search processes take place. Among them, RecA-mediated target search is the most thoroughly studied system. Qi *et al.* [**?**] selectively observed stable interactions between a RecA-ssDNA homologue and DNA in a DNA curtain experiment, in which single-molecule signals were only observed when ssDNA and dsDNA matched with each other for at least 8 nucleotides. Furthermore, using singlemolecule FRET, Ragunathan *et al.* [**?**] observed short-lived interactions (1-10 s) between RecA-ssDNA and target DNA that had 5-7 matching nucleotides. The difference between having 7 or 8 matches suggests there exists a rate limiting step hampering RecA-ssDNA filaments to extend base pairing beyond the 7th nucleotide (similar to the barrier representing the movement of the helix-7 motif in Figure **??**B).

Recently, great attention has been brought to the CRISPR/Cas system, an adaptive immune system in bacteria, which uses RNA as a guide to target foreign DNA or RNA [**?**]. CRISPR's target search involves a protein- DNA interaction (recognition of a 3-nt sequence, so-called PAM sequence) and RNA-DNA interactions. Biochemical studies suggested that it is the PAM recognition that occurs prior to the seed recognition [**? ? ?**]. Recently, a structural study showed that the first 8 nucleotides of Cas9's guide are pre-organized in a helical Aform, similar to the seed sequence of microRNA in Argonaute [**?**]. A recent FRET study indicated that there is another mode that follows binding of the seed recognition [**?**]. The authors showed that only when the guide RNA of Cas9 makes extensive base pairing (16nt out of the 20nt guide), a nuclease domain (HNH) migrates towards the target DNA. Altogether, the findings imply that CRISPR/Cas9, similar to Argonaute, uses more than two binding modes to overcome the speed-stability paradox ('PAM only' to 'PAM+seed' to 'cleavage competent'). Whereas a DNA curtain assay ruled out long distance lateral diffusion, it will be interesting to find out whether the CRISPR-Cas system makes any local lateral excursions when searching for the PAM sequence. Similarly, no large scale lateral diffusion has been observed for RecA/Rad51 systems using DNA curtain assays (>100nm resolution) [**?**], while short-range lateral diffusion was observed in single-molecule FRET experiments (nanometer resolution) [**?**] .

Finally, it will be interesting to find out how much the search mechanism of human Argonaute-2 is shared with other target search systems such as those mentioned in this review and different classes of Ago proteins that use DNA to target DNA [**? ?**] and RNA to target DNA [**?**] as well as PIWI proteins [**?**].

4.6. Acknowledgements

C.J. was funded by European Research Council under the European Union's Seventh Framework Programme [FP7/2007-2013] / ERC grant agreement n° [309509]. M.D. was supported by a TU Delft start up grant. This work was supported by the Netherlands Organization for Scientific Research (NWO/OCW), as part of the Frontiers in Nanoscience program. We thank Aafke van den Berg and Tao Ju (Thijs) Cui for critically reading and commenting on the manuscript.

References

[] D. P. Bartel, *MicroRNAs: Target Recognition and Regulatory Functions,* Cell **136**, 215 (2009), arXiv:0208024 [gr-qc] .

[] V. N. Kim, J. Han, and M. C. Siomi, *Biogenesis of small RNAs in animals,* Nature Reviews Molecular Cell Biology **10**, 126 (2009), arXiv:NIHMS150003 .

[] A. D. Riggs, S. Bourgeois, and M. Cohn, *The lac represser-operator interaction. III. Kinetic studies,* Journal of Molecular Biology **53**, 401 (1970).

[] O. G. Berg, R. B. Winter, and P. H. von Hippel, *Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory.* Biochemistry **20**, 6929 (1981).

[] P. H. Vonhippel and O. G. Berg, *Facilitated Target Location in Biological-Systems,* Journal of Biological Chemistry **264**, 675 (1989).

[] M. Bauer and R. Metzler, *Generalized facilitated diffusion model for DNA-binding proteins with search and recognition states,* Biophysical Journal **102**, 2321 (2012).

[] M. Coppey, O. Bénichou, R. Voituriez, and M. Moreau, *Kinetics of target localization of a protein on DNA: A stochastic approach,* Biophysical Journal **87**, 1640 (2004).

[] M. Slutsky and L. A. Mirny, *Kinetics of Protein-DNA Interaction : Facilitated Target Location in Sequence-Dependent Potential,* Biophysical Journal **87**, 4021 (2004).

[] H.-x. Zhou, *Rapid search for specific sites on DNA through conformational switch of nonspecifically bound proteins,* Proceedings of the National Academy of Sciences (2011), 10.1073/pnas.1101555108.

[] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf, *The lac Repressor Displays Facilitated,* Science **336**, 1595 (2012).

[] A. Tafvizi, F. Huang, A. R. Fersht, L. A. Mirny, and A. M. van Oijen, *A single-molecule characterization of p53 search on DNA,* Proceedings of the National Academy of Sciences **108**, 563 (2011).

[] P. C. Blainey, A. M. van Oijen, A. Banerjee, G. L. Verdine, and X. S. Xie, *A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA,* Proceedings of the National Academy of Sciences **103**, 5752 (2006).

[] J. Gorman, A. Chowdhury, J. A. Surtees, J. Shimada, D. R. Reichman, E. Alani, and E. C. Greene, *Dynamic Basis for One-Dimensional DNA Scanning by the Mismatch Repair Complex Msh2-Msh6,* Molecular Cell **28**, 359 (2007).

[] C. Jeong, W. K. Cho, K. M. Song, C. Cook, T. Y. Yoon, C. Ban, R. Fishel, and J. B. Lee, *MutS switches between two fundamentally distinct clamps during mismatch repair,* Nature Structural and Molecular Biology **18**, 379 (2011), arXiv:15334406 .

[] L. Zandarashvili, A. Esadze, D. Vuzman, C. A. Kemme, Y. Levy, and J. Iwahara, *Balancing between affinity and speed in target DNA search by zinc-finger proteins via modulation of dynamic conformational ensemble,* Proceedings of the National Academy of Sciences **112**, E5142 (2015).

[] K. Ragunathan, C. Liu, and T. Ha, *RecA filament sliding on DNA facilitates homology search,* eLife , 1 (2012).

[] Z. Chen, H. Yang, and N. P. Pavletich, *Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures,* Nature **453**, 489 (2008), arXiv:NIHMS150003 .

[] E. Elkayam, C. D. Kuhn, A. Tocilj, A. D. Haase, E. M. Greene, G. J. Hannon, and L. Joshua-Tor, *The structure of human argonaute-2 in complex with miR-20a,* Cell **150**, 100 (2012).

[] K. Nakanishi, D. E. Weinberg, D. P. Bartel, and D. J. Patel, *Structure of yeast Argonaute with guide RNA,* Nature **486**, 368 (2012).

[] N. T. Schirle and I. J. MacRae, *The crystal structure of human Argonaute2.* Science **336**, 1037 (2012), arXiv:NIHMS150003 .

[] Y. Wang, G. Sheng, S. Juranek, T. Tuschl, and D. J. Patel, *Structure of the guide-strand-containing argonaute silencing complex,* Nature **456**, 209 (2008), arXiv:15334406 .

[] S. D. Chandradoss, N. T. Schirle, M. Szczepaniak, I. J. Macrae, and C. Joo, *A Dynamic Search Process Underlies MicroRNA Targeting,* Cell **162**, 96 (2015).

[] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj, *How a protein searches for its site on DNA : the mechanism of facilitated diffusion,* Journal of Physics A: Mathematical and Theoretical **42**, 1 (2009).

[] U. Gerland, J. D. Moroz, and T. Hwa, *Physical constraints and functional characteristics of transcription factor-DNA interaction,* Proceedings of the National Academy of Sciences **99**, 12015 (2002), arXiv:0112083 [physics] .

[] R. Zwanzig, *Diffusion in a rough potential,* Proceedings of the National Academy of Sciences **85**, 2029 (1988).

[] O. Benichou, Y. Kafri, M. Sheinman, and R. Voituriez, *Searching Fast for a Target on DNA without Falling to Traps,* Physical Review Letters **138102**, 1 (2009).

[] J. Reingruber and D. Holcman, *Transcription factor search for a DNA promoter in a three-state model,* Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **020901**, 1 (2011).

[] R. Murugan, *Theory of Site-Specific DNA-Protein Interactions in the Presence of Conformational Fluctuations of DNA Binding Domains,* Biophysical Journal **99**, 353 (2010).

[] S. Yu, S. Wang, and R. G. Larson, *Proteins searching for their target on DNA by one-dimensional diffusion : overcoming the " speed-stability " paradox,* Journal of Biological Physics , 565 (2013).

[] S. L. Ameres, J. Martinez, and R. Schroeder, *Molecular Basis for Target RNA Recognition and Cleavage by Human RISC,* Cell **130**, 101 (2007).

[] M. Khorshid, J. Hausser, M. Zavolan, and E. Van Nimwegen, *A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets,* Nature Methods **10**, 253 (2013).

[] L. M. Wee, C. F. Flores-Jasso, W. E. Salomon, and P. D. Zamore, *Argonaute divides Its RNA guide into domains with distinct functions and RNA-binding properties,* Cell **151**, 1055 (2012), arXiv:NIHMS150003 .

[] N. T. Schirle, J. Sheu-Gruttadauria, and I. J. Macrae, *Structural basis for microRNA targeting,* Science **346** (2014), 10.1126/science.1258040.

[] W. E. E. Salomon, S. M. M. Jolly, M. J. J. Moore, P. D. D. Zamore, and V. Serebrov, *Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides,* Cell **162**, 84 (2015).

[] M. H. Jo, S. Shin, S. R. Jung, E. Kim, J. J. Song, and S. Hohng, *Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs,* Molecular Cell **59**, 117 (2015).

[] D. Herschlag, C. M. Gherghe, and K. M. Weeks, *Implications of ribozyme kinetics for targeting the cleavage of specific RNA molecules in vivo: more isn't always better.* Proceedings of the National Academy of Sciences of the United States of America **88**, 6921 (1991).

[] S. R. Jung, E. Kim, W. Hwang, S. Shin, J. J. Song, and S. Hohng, *Dynamic anchoring of the 3�-end of the guide strand controls the target dissociation of argonaute-guide complex,* Journal of the American Chemical Society **135**, 16865 (2013).

[] A. Deerberg, S. Willkomm, and T. Restle, *Minimal mechanistic model of siRNA-dependent target RNA slicing by recombinant human Argonaute 2 protein,* Proceedings of the National Academy of Sciences **110**, 17850 (2013).

[] T. Kawamata and Y. Tomari, *Making RISC,* Trends in Biochemical Sciences **35**, 368 (2010).

[] H. M. Sasaki and Y. Tomari, *The true core of RNA silencing revealed,* Nature Structural and Molecular Biology **19**, 657 (2012).

[] A. Zander, P. Holzmeister, D. Klose, P. Tinnefeld, and D. Grohmann, *Single-molecule FRET supports the two-state model of argonaute action,* RNA Biology **11**, 45 (2014).

[] N. T. Schirle, J. Sheu-Gruttadauria, S. D. Chandradoss, C. Joo, and I. J. MacRae, *Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets,* eLife **4**, 1 (2015), arXiv:arXiv:1011.1669v3 .

[] S. H. Sternberg, B. Lafrance, M. Kaplan, and J. A. Doudna, *Conformational control of DNA target cleavage by CRISPR-Cas9,* Nature **527**, 110 (2015).

[] M. Thomas, J. Lieberman, and A. Lal, *Desperately seeking microRNA targets,* Nature Structural and Molecular Biology **17**, 1169 (2010).

[] D. P. Bartel, *TargetScan,* (2015).

[] V. Agarwal, G. W. Bell, J. W. Nam, and D. P. Bartel, *Predicting effective microRNA target sites in mammalian mRNAs,* eLife **4**, 1 (2015).

[] J. A. Broderick, W. E. Salomon, S. P. Ryder, N. Aronin, and P. D. Zamore, *Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing,* Rna **17**, 1858 (2011).

[] A. Grimson, K. K. H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel, *MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing,* Molecular Cell **27**, 91 (2007).

[] P. Sætrom, B. S. E. Heale, O. Snøve, L. Aagaard, J. Alluin, and J. J. Rossi, *Distance constraints between microRNA target sites dictate efficacy and cooperativity,* Nucleic Acids Research **35**, 2333 (2007).

[] Z. Qi, S. Redding, P. Sung, E. C. Greene, Z. Qi, S. Redding, J. Y. Lee, B. Gibb, Y. Kwon, H. Niu, W. A. Gaines, and P. Sung, *DNA Sequence Alignment by Microhomology Sampling during Homologous Recombination Article DNA Sequence Alignment by Microhomology Sampling during Homologous Recombination,* Cell **160**, 856 (2015).

[] B. Wiedenheft, S. H. Sternberg, and J. A. Doudna, *RNA-guided genetic silencing systems in bacteria and archaea,* Nature **482**, 331 (2012), arXiv:37 .

[] E. Semenova, M. M. Jore, K. A. Datsenko, A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. J. Brouns, and K. Severinov, *Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence,* Proceedings of the National Academy of Sciences **108**, 10098 (2011).

[] S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna, *DNA interrogation by the CRISPR RNA-guided endonuclease Cas9,* Nature **507**, 62 (2014), arXiv:NIHMS150003 .

[] E. R. Westra, E. Semenova, K. A. Datsenko, R. N. Jackson, B. Wiedenheft, K. Severinov, and S. J. Brouns, *Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base*

Pairing-Independent PAM Recognition, PLoS Genetics **9** (2013), 10.1371/journal.pgen.1003742.

[] F. Jiang, K. Zhou, S. Gressel, and J. A. Doudna, *A cas9 guide RNA complex preorganized for target DNA recognition,* Science **348**, 1477 (2015).

[] A. Graneli, C. C. Yeykal, R. B. Robertson, and E. C. Greene, *Long-distance lateral diffusion of human Rad51 on double-stranded DNA,* Proceedings of the National Academy of Sciences **103**, 1221 (2006).

[] F. Gao, X. Z. Shen, F. Jiang, Y. Wu, and C. Han, *DNA-guided genome editing using the Natronobacterium gregoryi Argonaute,* Nature Biotechnology **34**, 768 (2016).

[] D. C. Swarts, M. M. Jore, E. R. Westra, Y. Zhu, J. H. Janssen, A. P. Snijders, Y. Wang, D. J. Patel, J. Berenguer, S. J. Brouns, and J. Van Der Oost, *DNA-guided DNA interference by a prokaryotic Argonaute,* Nature **507**, 258 (2014), arXiv:15334406 .

[] I. Olovnikov, K. Chan, R. Sachidanandam, D. K. Newman, and A. A. Aravin, *Bacterial Argonaute Samples the Transcriptome to Identify Foreign DNA,* Molecular Cell **51**, 594 (2013), arXiv:NIHMS150003 .

[] R. J. Ross, M. M. Weiner, and H. Lin, *PIWI proteins and PIWI-interacting RNAs in the soma,* Nature **505**, 353 (2014), arXiv:NIHMS150003 .

5

Argonaute bypasses cellular obstacles without hindrance during target search

Argonaute (Ago) proteins are key players in both gene regulation (eukaryotes) and host defense (prokaryotes). Acting on single-stranded nucleic-acid substrates, Ago relies on base pairing between a small nucleic-acid guide and its complementary target sequences for specificity. To efficiently scan nucleic-acid chains for targets, Ago diffuses laterally along the substrate and must bypass secondary structures as well as protein barriers. Using single-molecule FRET in conjunction with kinetic modelling, we reveal that target scanning is mediated through loose protein-nucleic acid interactions, allowing Ago to slide short distances over secondary structures, as well as to bypass protein barriers via intersegmental jumps. Our combined single-molecule experiment and kinetic modelling approach may serve as a novel platform to dissect search process and study the effect of sequence on search kinetics for other nucleic acid-guided proteins.

This chapter has been published as: Cui, T.J., **Klein, M.**, Hegge, J.W., Chandradoss, S.D., van der Oost, J., Depken, M., and Joo, C. *Argonaute bypasses cellular obstacles without hindrance during target search.* Nature Communications **10**-4390 (2019)

5.1. Introduction

Target recognition by oligonucleotide guides is essential in cellular development, differentiation and immunity [**?** **?**]. Argonaute (Ago) proteins are key mediators of the target interference process, utilizing short oligo-nucleotides (20-30 nt) as guides for finding complementary target sequences [**?** **?**]. The guide-target interaction initiates at the 5' end of the guide, and progresses through Watson-Crick base pairing at the "seed" segment, which propagates along the guide, resulting in target interference upon completion [**?**]. While eukaryotic Argonautes use RNA guides to target RNA, prokaryotic Agos (pAgo) have been demonstrated to use a variety of guides and targets [**?** **?** **?**]. Depending on the pAgo type, it uses either DNA or RNA guides to target single-stranded (ss) DNA, ssRNA or both2. The ability of pAgos to cleave ssDNA but not double stranded DNA (dsDNA) suggests a physiological role as a host defense system against ss mobile genetic elements6–8. Recently, a new family of CRISPR-Cas systems that targets ssDNA—not dsDNA—have been discovered in archaea, suggesting that these defense systems may be more widespread than previously thought [**?**]. The number of potential targets encoded in cellular DNA/RNA is vast [**?** **?** **?**] and Ago needs to search long stretches of polymer before finding a canonical target. Single-molecule studies have shown that a mixture of excursions into solution and one-dimensional movements results in a search that is orders of magnitude more efficient than is possible without lateral diffusion [**?** **?**]. In a previous biophysical study we suggested that human Argonaute 2 (hAGO2) uses lateral diffusion along RNA for target search [**?**]. Yet, the degree of lateral diffusion remains unclear, as excessive usage of 1D diffusion would lead to redundant re-sampling of potential target sites and to problems at various roadblocks present on the target nucleic acids [**?** **?**]. In addition to complete dissociation into solution, intersegmental jumping, in which a protein transfers between two spatially close-by segments, has been shown to occur for DNA binding proteins such as restriction enzyme EcoRV [**?**]. After binding to DNA non-specifically from solution, the protein diffusively scans only a limited section [**?** **?** **?** **?**], and dissociates into solution before rebinding to a new section. Use of such a mechanism would lead to reduced sampling redundancy, and the possibility to circumvent obstructions when proteins search for their targets.

Previous studies have shown that certain DNA/RNA-guided proteins interact with DNA through non-specific electrostatic interactions [**?** **?** **?**], but the strength of these interactions and their behaviour on roadblocks and secondary structures is not understood. Since these interactions are typically short-ranged [**?** **?** **?**] and short-lived [**?** **?** **?** **?** **?** **?** **?**], a method offering high spatio-temporal resolution is required to study these interactions. Here we make use of single molecule Förster Resonance Energy Transfer (FRET) to elucidate the mechanism of ssDNA target search by a mesophilic Ago from the bacterium *Clostridium butyricum* (CbAgo). We show that CbAgo does not remain in tight contact with the DNA backbone, enabling it to bypass secondary structures along the nucleic-acid chain—all while retaining the ability to recognize its target. After sliding locally, the protein is able to reach distant sites (>100 nt) along the DNA through intersegmental jumps and then resumes sliding. These different modes of facilitated diffusion allow Ago to rapidly search through nucleic acid segments, as well as to bypass substantial obstacles during target scanning.

5.2. Results

5.2.1. Single-molecule kinetics of CbAgo binding

To elucidate the complexity of the target search mechanism, we made use of the high spatial sensitivity of single-molecule FRET. We studied a minimal Argonaute complex that consists of CbAgo, loaded with a 22-nt DNA guide (small interfering DNA, siDNA) [**?**]. By using total internal reflection fluorescence (TIRF) microscopy, we recorded the interactions of CbAgo-siDNA with target DNA. Target DNA was immobilized on a PEG-coated quartz surface in a microfluidic chamber through biotin-streptavidin conjugation. Guide-loaded CbAgo was introduced to the microfluidic chamber by flow. The target was embedded within a poly-thymine sequence and labelled with an acceptor dye (Cy5) (**Figure ??**a). The guide construct was labelled at nt 9 from the 5'-end with a donor dye (Cy3) (**Figure ??**b). A 532-nm laser excitation resulted in donor excitation when the protein loaded with the guide DNA interacted with the target DNA. Once the CbAgo-siDNA complex became bound to the target, the proximity of the donor dye to the acceptor dye on the target resulted in high FRET efficiency. This was followed by a sudden disappearance of the signal, indicating that the complex dissociated from the target and diffused into the free solution. Freely diffusing molecules move too rapidly ($\sim \mu$s) in and out of the evanescent field for the current time resolution of the experimental setup (100 ms) and were therefore not recorded. We found that CbAgo is not able to target dsDNA directly (**Figure ??**a-b). Likewise, when a ssDNA target with one base pair complementarity to the seed motif of the guide was used, only transient interactions (~0.45 s) were detected (**Figure ??**c-d), and no accurate binding profile could be extracted from the FRET histogram (**Figure ??**e). To observe target search that involves intrinsically transient interactions, we determined the optimal target motif for recording binding events. The optimal motif should provide binding events longer than our detection limit of 100 ms, but still lead to dissociation events within the time of our measurement (200 s). To determine the optimal motif, the complementarity between guide and target was incrementally extended from nt 2 to 8 of the guide, showing a gradually increasing dwell time of the Ago-siDNA complex. We found that increasing the number of complementary base pairs above 6 resulted in stable binding beyond the photobleaching time (**Figure ??**c). To maintain weak interactions, we continued our experiments using a siDNA with three-base complementarity (N=3) with the target (nt 2-4) (**Figure ??**f). This gives a well-defined FRET population in the FRET histogram (**Figure ??**h), unlike one base-pair complementarity. Our estimation of the photobleaching rate (1.4 x 10^{-3} s^{-1}) (**Figure ??**d) was an order of magnitude lower than the dissociation rate (2.7 x 10^{-2} s^{-1}) (**Figure ??**g), indicating that photobleaching does not affect our estimation of the dissociation rate.

5.2.2. Lateral diffusion of CbAgo

It was previously shown that an Ago-guide complex does not directly bind a specific target site from solution, but rather binds non-specifically to random positions along a surfaced-immobilized nucleic acid construct [**?**]. Such non-specific interactions of CbAgo-siDNA along target DNA are too short-lived to resolve in the absence of a canonical target motif (**Figure ??**c), and in the presence of such a motif there was still no lateral diffusion visible (**Figure ??**f). As we were unable to resolve lateral diffusion by CbAgo from non-specifically bound regions to the target, we questioned whether the observed stable signal for three

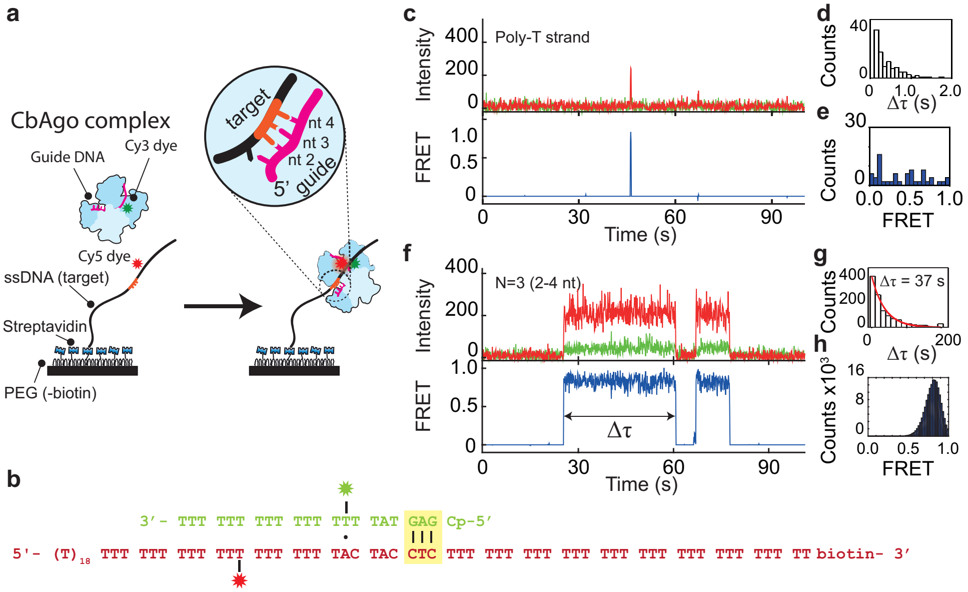


Figure 5.1: Single molecule imaging of target binding by siDNA:CbAgo complex. a, Immobilization scheme of the Argonaute-guide DNA complex. ssDNA is immobilized on a PEGylated quartz slide surface. Presence of the Ago-siDNA complex is detected by specific binding to target site (light yellow) resulting in high FRET. b, Sequences of guide (green) and target DNA. Guide is labelled on the 9th nucleotide position from the 5' side. c, Representative FRET trace of a single molecule experiment at 100 mM NaCl showing a transient interaction between CbAgo and a poly-T strand. Time resolution is 100 ms. d, Dwell time distribution of the Argonaute in absence of target motif. e, FRET values of the transient interactions of (d). f, Representative FRET trace of a single molecule experiment showing the interaction between CbAgo and a 2-4 nt (N=3) motif. g, Dwell time distribution Dwell time distribution of N=3 binding events with the mean dwell time of 37 s. h, FRET histogram of binding events, showing a single FRET population for N=3 at E=0.78.

complementary base pairs is due to stable binding to the target or contains lateral excursions away from the target but below our time resolution. In case of the latter, measured apparent dwell times (**Figure ??**g) would consist of the combined dwell times of many target escapes through lateral diffusion, each followed by rapid recapture below the detection limit, before CbAgo eventually unbinds from the DNA (**Figure ??**g). We show that such a process of repeated recapture would result in an exponential distribution of apparent dwell times, in accordance with **Figure ??**g (see **S.I.**). To overcome the temporal resolution limit, we adopted a tandem target assay [**? ?**]. While lateral diffusive excursions from a trap are too short-lived to be resolved in the presence of only a single target, a second target can trap an excursion for long enough to be observed. We placed two identical optimal targets (site 1 and site 2) separated by 22 nt (**Figure ??**a) along the DNA construct. Both targets base pair only with the first three nucleotides (nt 2-4) of the guide bound by CbAgo. As the second target is located further away from the acceptor dye, binding the second target results in a lower FRET efficiency than binding the first target. This difference in FRET values allows us to determine which of the two targets CbAgo-siDNA is bound to (**Figure ??**b). The respective distance and FRET efficiency between the first binding site (site 1) and the acceptor dye (Cy5) remained the same as for the single target assay (E 0.78), while an additional peak appeared at a lower FRET efficiency for the second target (E 0.43, **Figure ??**c). After binding to one of the target sites, a majority of the binding events (87.8%) resulted in CbAgo-siDNA shuttling to the other target without loss of FRET signal. Under our standard experimental condition (100 mM NaCl), an average of 13.5 shuttling events occur per binding event (**Figure ??**d). When the experiment was repeated for guides and targets with complementary increased to N=6 (nt 2-7), only 15.1% of the traces showed the shuttling signature within our time window (**Figure ??**f). This shows that the shuttling signature is controlled by interactions between CbAgo-ssDNA and the target motif. With a 6-nt match, the target is strongly bound, and we are less likely to observe a shuttling event within our observation window.

Interestingly, the average dwell time of the first target (**Figure ??**g) decreased from 37 s to 1.7 and 1.8 s after adding a second target in its vicinity (**Figure ??**e). This observation is in agreement with our lateral diffusion model, since with close-by targets, each sub-resolution diffusive excursion has some probability to be caught at the opposing target. To further test our claim that the transition between targets occur through lateral diffusion, we use single-molecule analysis software [**?**] to extract the average time between shuttling events ($\Delta\tau_{\text{shuttle}}$) from traces (**Figure ??**).

5.2.3. Kinetic modelling of lateral diffusion

To determine how lateral diffusion contributes to the shuttling, we kinetically model how $\Delta\tau_{\text{shuttle}}$ depends on the distance between traps. The DNA construct is modelled as a series of binding sites along which CbAgo will perform an unbiased random walk by stepping to neighboring nucleotides. The rate of stepping away from the target is k_{esc} in both directions, while at non-specific sites (poly-T), stepping is assumed to be near instantaneous—an approximation justified by the fact that lateral excursions are never resolved in the experiments. The time needed for FRET transitions to occur (named “shuttling time”, $\Delta\tau_{\text{shuttle}}$) is equivalent to the apparent dwell time at a single FRET state. In the **S.I.** we construct a diffusive model capturing the effect of Ago’s repeated retrapping before shuttling to the other trap. The model shows that the shuttling time from the target grows linearly with the separation x_{target} between the targets

$$\Delta\tau_{\text{shuttle}}(x_{\text{target}}) = \frac{x_{\text{target}}}{k_{\text{esc}}} \tag{5.1}$$

The linear dependence of the shuttling time with trap separation might seem puzzling at first, given that diffusive timescales usually show a quadratic dependence on distances. Here though, it is not the diffusive steps themselves that directly contributes to the shuttling time, but rather the changing probability to getting retrapped before shuttling. In support of this model, we observed that the apparent shuttling time $\Delta\tau_{\text{shuttle}}(x_{\text{target}})$ increases approximately linearly when the distance between the targets increases through 11, 15, 18 and 22 nt (**Figure ??**). A fit to Equation 1 reveals that CbAgo-siDNA complexes escape the target site at a rate of 15.8 times per second ($k_{\text{esc}} = 15.8s^{-1}$) in either direction.

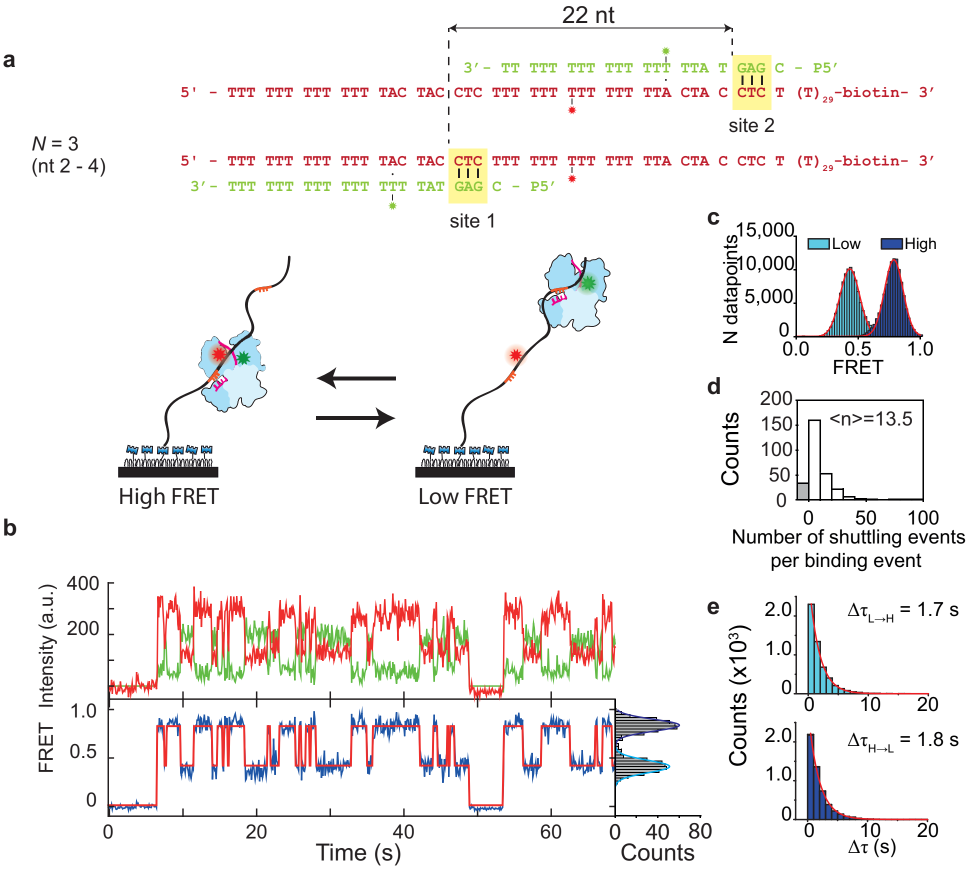


Figure 5.2: Shuttling signature of CbAgo appears in presence of two targets. a, In the top right corner the DNA sequence of guide and target for 22 nt separation between targets. Here the distance is defined as the distance from beginning of a target to the beginning of the next target. The placement of the second target (site 2) results in the appearance of an additional FRET signal, with lower FRET efficiency. b, (Top) Representative shuttling trace of a 22 nt separation tandem target at 100 mM NaCl for N=3. (Bottom) The corresponding FRET states (blue) with the fitted HMM trace on top (red). (Right) FRET histogram of the respective time trace. Time resolution is 100 ms. c, FRET histograms of respective states, with peaks at 0.43 and 0.78. d, Shuttling event distribution for the same conditions (n=309). Bin size = 10. On average 13.5 shuttling events take place before dissociation. The grey bar (n=33) marks binding events followed by dissociation (no shuttling). e, Dwell time distributions of respectively the transitions from low FRET state to high FRET state (top) and vice versa (bottom).

5.2.4. Ago probes for targets during lateral diffusion

Next, we placed a third target on the tandem construct (**Figure ??**a), keeping the distance between each set of neighboring targets well within the regime for which we find good agreement to Equation 1 using the assay discussed above (i.e. at 11 nt trap separation, see **Figure ??**). We observed three different FRET levels, corresponding to CbAgo getting trapped at the three different targets (**Figure ??**b). Using Hidden Markov Modelling (HMM), states can be assigned (**Figure ??**b) and transition probabilities can be extracted (**Figure ??**c). If CbAgo returns to solution between binding targets, transitions between

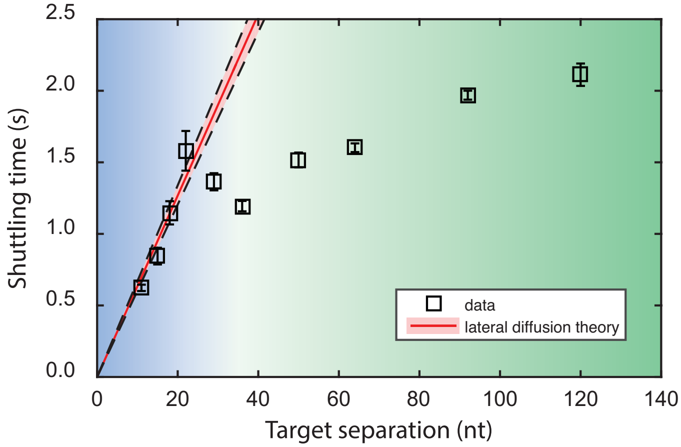


Figure 5.3: CbAgo shuttling behaviour differs across short and large distances Shuttling time is plotted versus distance between targets. Squares indicate the mean shuttling time for each DNA construct. The plotted error bars indicate the 95% confidence interval of 10^5 bootstrapped dwell times. The red line indicates the lateral diffusion model where the first four data points are fitted with $k_{\text{esc}} = 15.8s^{-1}$. The shaded red region indicates its 95% confidence interval. The blue region indicates where the shuttling time follows lateral diffusion theory. This theory breaks down for larger distances (green).

any pair of targets will be equally probable, resulting in equal effective rates between all targets. However, if lateral diffusion dominates, transitions between adjacent sites will be favored. The transition probabilities (**Figure ??**c) indicate that over 90% of the transitions between the two outer targets (from state A to C, or from C to A) proceed through the intermediate target site (state B). The rate to transfer from B to C and B to A is twice as much as that of the opposite path (A to B or C to B). Using the fitted escape rate from above, $k_{\text{esc}} = 15.8s^{-1}$, we predict similar shuttling times based on our theoretical model for lateral diffusion (**Figure ??**d, **S.I.**). With no more free-parameters remaining for this prediction, we take this experimental agreement with our prediction as further evidence of lateral diffusion. It is noteworthy that there are about 10% direct transitions from A to C and C to A without any intervening dissociation. The exponential distribution of the dwell times (**Figure ??**b) suggests that at our current time resolution this 10% may be either due to missed events or due to the existence of an additional translocation mode through which Ago is able to bypass the intermediate target.

5.2.5. Ago target search is unhindered by structural and protein barriers

Secondary structures are commonly found in mRNA and are also predicted to exist in single stranded viruses [**? ?**]. It is not known whether CbAgo is able to bypass the numerous junctions it encounters upon scanning a DNA segment. To examine this, a Y-fork structure (DNA junction) was introduced as a road block between two targets (**Figure ??**a), while keeping their separation (11 nt) the same as in the tandem target variant (**Figure ??**f). The construct was designed such that the labelled target was partially annealed at the stem with a biotinylated target, thus only annealed constructs were observable on the surface

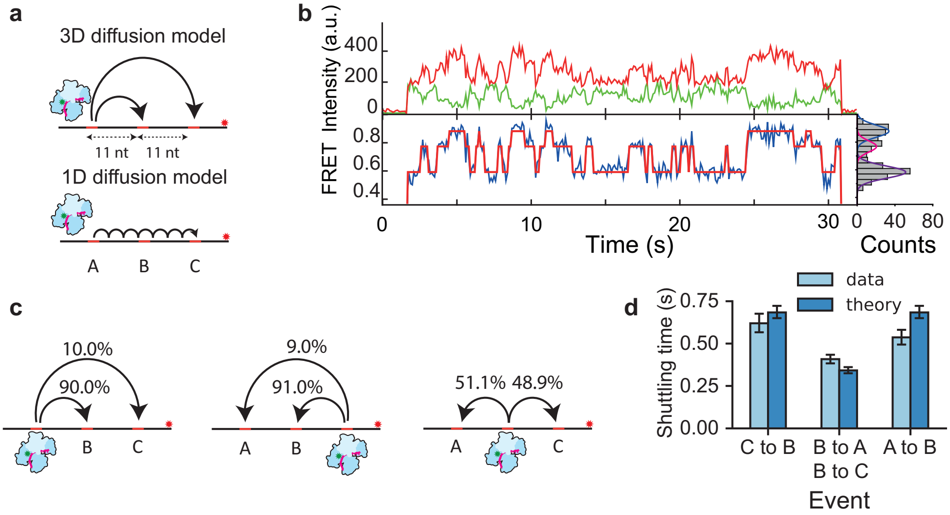


Figure 5.4: CbAgo undergoes short range diffusion through correlated steps. a, Models for target translocation at short range. In the 3D diffusion model, target dissociation occurs from A followed by random 3D diffusion through solution. In effect, the neighboring two targets (B and C) will compete for binding. In the lateral diffusion model, the CbAgo complex will have to bypass the adjacent target B before binding to target C. b, Representative FRET trace showing the shuttling behavior between three targets. Top: donor (green) and acceptor (red) intensities. Bottom: FRET trace (blue) and HMM assigned states (red). Right: The fitted states from this data trace with dark blue: state C, pink: state B and purple state A. c, Transition probabilities from state A to B,C (left), from state C to A and B (middle) and from state B to A or C (right). d, Experimental values of the shuttling time of the three target construct were compared against the parameter-free theoretical model that only uses the $k_{\text{esc}} = 15.8s^{-1}$ from Figure 3. Error bars indicate the 95% confidence interval acquired from 10^5 bootstraps.

of the microfluidic device. When CbAgo binds to either of the two targets, it can reach the other target only by crossing the junction. Our measurement showed that there was no significant difference in shuttling time between the standard tandem-target construct and the Y-fork construct (**Figure ??**b-c), indicating that the Y-fork does not impede any of the lateral diffusion modes present. We have previously observed that the CbAgo-siDNA complex is not able to stably bind to dsDNA31, demonstrating that the protein cannot simply track the backbone of dsDNA (**Figure ??**a-b). Thus, our result suggests that the Ago-siDNA complex does not maintain tight contact with DNA during lateral diffusion. Maintaining a weak interaction with the DNA molecule allows CbAgo-siDNA to move past the junction. Next, we questioned whether CbAgo is also able to overcome larger barriers, such as proteins which cannot reasonably be traversable through sliding alone. Lin28, a sequence-specific inhibitor of let-7 miRNA biogenesis, has been found to associate sequence specifically to RNA and DNA [**?**]. His-tagged Lin28 was immobilized on the surface of the microfluidic chamber (**Figure ??**d) after which a fluorescent ssDNA fragment was added containing a central Lin28 binding motif and an Ago target motif on either side (**Figure ??**d & **Figure ??**g). The presence of the protein blockade did not preclude Ago from reaching the distal site (**Figure ??**e) but noticeably broadened the FRET peak (**Figure ??**f), possibly due to protein-protein interactions. Although the shuttling rate was lowered from $0.60s^{-1}$ to $0.27s^{-1}$ (**Figure ??**g & **Figure ??**e), Ago is able to bypass the obstacle. Since short-range lateral movement is now blocked by the protein barrier, Ago's ability to move between targets demonstrates that the target search process also allows for intersegmental jumps, in accordance with our observation that the middle target is sometimes skipped when transitioning between the outer targets in **Figure ??**c.

5.2.6. Ago relies on flexibility of DNA segments of bypassing blockades

Since Ago was observed to be able to bypass junctions and proteins, we questioned whether Ago could bypass other large-profile barriers. Previously, we observed that Ago only interacts transiently with dsDNA (**Figure ??**a-b) and thus we repurposed dsDNA as an extended blockade. We made a construct analogous to the tandem target construct used in **Figure ??**a, but the targets were separated by 36 nt and complementary strands of 17 nt, 21 nt, and 25 nt were annealed to the region in between the targets (**Figure ??**h-i). For the construct with a 17-nt blockade we observed a large number of shuttling events (shuttling probability 65.3% upon binding) indicating that a dsDNA blockade does not prohibit CbAgo from reaching the other site (**Figure ??**j and **Figure ??**l black squares). Upon extending the length of the dsDNA blockade, to 21 nt and 25 nt, we noticed a drop in the percentage of shuttling events (63.1% and 40.4% respectively) although shuttling still persisted (Supplementary Fig **??**). Since the stiff segment of dsDNA decreases the shuttling probability, we conclude that Ago relies on the flexibility of segments for lateral diffusion. To further investigate the contribution of DNA flexibility, we used another construct which was shortened (by 15 nt from 19 nt) from the 5' side (**Figure ??**h bottom sequence). Here, ssDNA coiling was no longer possible from the 5' side of the DNA construct (**Figure ??**k). We measured a significant decrease (50%) in shuttling probability for all three blockades compared to the untruncated construct (**Figure ??**l), which supports that Ago relies on the flexibility of DNA segments when transferring between them.

5.2.7. Ago uses hops to access distant DNA segments

Sliding is not expected to dominate across large distances, as the linear increase in shuttling time (Equation **??**) would render the search process prohibitively slow. However, when CbAgo was studied with tandem targets that were separated 36 nt or more, we observed that the shuttling still persisted across larger distances (**Figure ??**, green region, Supplementary Table 1 and **Figure ??**). Together with the evidence of intersegmental jumping above, and the fact that the ssDNA can easily be coiled back to bring the second target close to the Ago protein [**?**], we speculate that there is a second mechanism of lateral diffusion: after local scanning for the target through sliding, the CbAgo complex jumps to a different part of the segment that has looped back into proximity of the complex. From this point on, we refer to these hops as intersegmental transfers in accordance with the current literature (**Figure ??**) [**? ?**]. This intersegmental jumping mechanism would enable CbAgo to travel to new sites without fully dissociating, and rescanning of the same sections would be minimized [**? ?**]. Based on the dependence of the single-target off-rate on the ionic strength (**Figure ??**f), we expect the rate of the intersegmental jumps to also be dependent on salt concentration, while sliding should only be moderately effected since it has no net effect on the ion condensation along the substrate. In order to test the hypothesis that short-ranged lateral diffusion is governed by sliding and long-range diffusion is governed

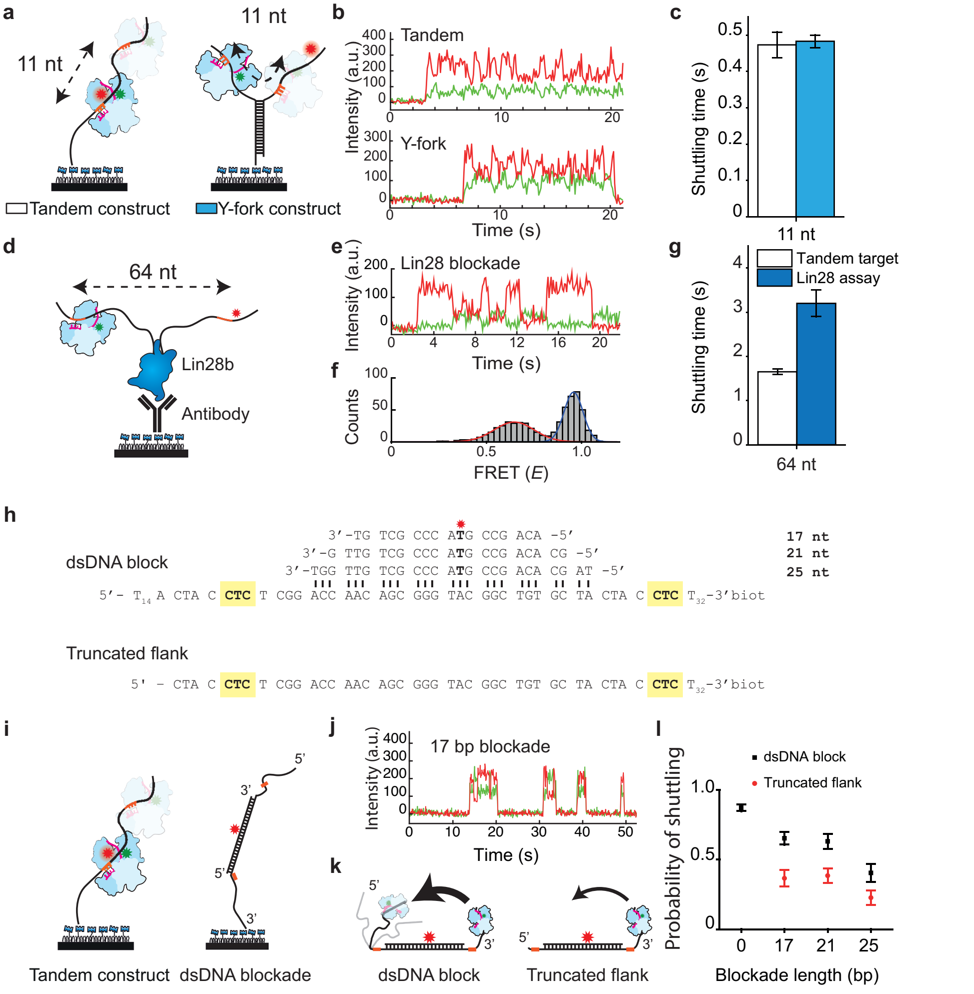


Figure 5.5: Argonaute can overcome structural and protein barriers. a, Schematic drawing tandem target assay (left) and the Y-fork assay (right) with 11 nt separation between targets. b, Representative shuttling traces of the tandem target assay (top) and Y-fork assay. c, The shuttling time of the Y-fork junction (blue bar) compared with the tandem assay (white bar). The experimental data of both sets were taken on the same days. Error bars indicate the 95% confidence interval acquired from 105 bootstraps d, Schematic drawing of the His-Lin28b blockade assay, where targets are separated by 64 nt. Immobilization happens through a biotin-anti-His antibody. e, Example of a shuttling trace with Lin28b located in between two targets. Exposure time is 100 ms. f, FRET histogram (molecules n = 46) fit with two Gaussian functions (E=0.64 for red fit and E=0.95 for dark blue fit). g, The shuttling time of the Lin28 assay compared with the tandem target assay for 64 nt separation between targets. h, Sequences used for the dsDNA block assay, indicating the base pairing between a 17 nt, 21 nt and a 25 nt long blockade and the target strand. The dsDNA block construct has a 19 nt flank on the 5' side, whereas the "truncated flank" has a 4 nt flank. i, Schematic of a dsDNA block assay, where the CTC targets are highlighted with orange. j, Representative trace of binding and shuttling of CbAgo on a 17 bp blockade DNA construct. k, (left) Schematic of dsDNA block construct with full length flanks. (right) schematic of the truncated version where the flank on the 5' side is removed. The thickness of the arrows indicate the observed shuttling probability. l, The probability of shuttling upon binding to a CTC target plotted versus the blockade length (none, 17 nt, 21 nt and 25 nt) for full length flanks (black squares) and for the truncated flanks (red circles). Error bars are given by the 95% confidence interval acquired from 10^5 bootstraps

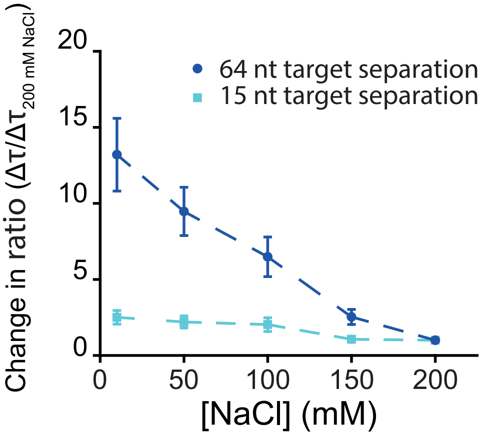


Figure 5.6: The relative change in shuttling time of two constructs from **Figure ??**, 64 nt separation (dark blue circles) and 15 nt separation (light blue squares), normalized against $\Delta\tau_{shuttle}$ at 200 mM NaCl. Errors of the ratio were determined through bootstrapping 10^5 times the ratio of $\Delta\tau/\Delta\tau_{200\text{ mM NaCl}}$

by intersegmental jumps, we altered the ionic strength of the buffer solution from 10 mM NaCl to 200 mM NaCl. Here, we expect the degree of DNA coiling not to be significantly affected by the change in salt concentration, since the persistence length is only expected to vary between 20 and 14 when exchanging the buffers, and in both buffers it is smaller than the contour length of the constructs [**?**]. We used dual-target constructs with 15-nt separation and 64-nt separation (**Figure ??**), taken from the two different regions in **Figure ??** (indicated by blue and green shading). At a separation of 64 nt, we observed a 13-fold increase of the shuttling rate when increasing the salt concentration from 10 mM NaCl to 200 mM NaCl. In contrast, we observed that for the dual-target construct with 15-nt separation, the shuttling time changed roughly only two-fold for the same change in ionic strength (**Figure ??**)—a modest change compared to 13-fold of the dual-target constructs with 64-nt separation. We take the relative ionic-strength insensitivity of shuttling times for 15-nt trap separation as evidence of translocation being dominate by sliding over short distances. In contrast, given the relative ionic-strength sensitivity for the 64-nt construct, the Ago complex is here unlikely to first reach the distal site through sliding only, and requires partial dissociation from the DNA strand. In conclusion, lateral diffusion during CbAgo target search is governed by two distinct modes. For short distances, lateral diffusion takes place through a sliding process characterized by loose contact with the DNA strand. This allows the protein to “glide” past secondary structures. To traverse larger distances, CbAgo is able to take advantage of the fact that the softness of the substrate allows it to bend back on itself to enable frequent intersegmental jumps between nearby segments (**Figure ??**).

5.3. Discussion

Within a vast number of potential targets, Ago-guide complexes have to minimize the time spent unproductively diffusing through solution or redundantly checking off-targets, as timely regulation is crucial for both cell development and host defense [**?**]. Our single-molecule study shows that Argonaute from C. butryicum (CbAgo) uses a loose sliding mode to bypass junctions and relies on intersegmental jumps to cover larger distances and to bypass substantial barriers.

We have shown that bacterial Ago binds DNA loosely and slides along the DNA to locally scan for complementary targets. While such sliding mechanism has been characterized for several proteins [**? ? ? ?**], little was previously known for DNA/RNA-guided target searchers like Ago. Proteins searching along nucleic acids with secondary structures may be blocked from sliding further. However, this does not seem to be true for Ago. Instead, the loose interaction with the substrate allows the protein to slide past junctions while still probing potential target sequence through base pairing. To the best of our knowledge, this mode of loose-contact sliding has not been reported for any nucleic-acid guided proteins. In addition, we show that the loose binding further allows Ago to move to a new segment via intersegmental jumps, reducing redundant scanning of the same segment and allowing Ago to bypass large-profile roadblocks.

The ability of CbAgo to target specifically ssDNA but not dsDNA [**?**] (**Figure ??**a-b) suggests a role as host defense against mobile genetic elements and ssDNA viruses. In environments where ssDNA viruses can be abundant, such as in sea water, fresh water, sediment, terrestrial, extreme, metazoan-associated and marine microbial mats [**? ? ?**], pAgo's targeting ssDNA would be very beneficial for the host. Upon entry in the infected cell, ssDNA binding and recombination proteins may associate with the invading nucleic acid, and DNA polymerase will start to generate the second strand. In addition, it is anticipated that secondary structures will be formed in the ssDNA viral genome [**?**]. This will generate road blocks that may affect scanning by defense systems such as restriction enzymes but—as shown here—not Argonaute. Likewise, insertion of transposons in prokaryotes often proceeds via a ssDNA-intermediate state [**? ? ?**], and pAgos may here encounter the same type of obstacles. In case of ssRNA, both in prokaryotes and in eukaryotes, it is well known that complex secondary structures can be formed by base pairing different anti-parallel RNA segments [**? ? ? ?**]. The presence of secondary structures suggests that it is necessary for Agos to "glide"—the type of loosely bound sliding we report—past such roadblocks to enable search along ssRNA. Based on the functional and structural similarities of prokaryotic Agos and eukaryotic Agos [**? ?**], we expect eAgo to also slide past RNA secondary structures, minimizing time spent trapped at such structures.

The effect of lateral diffusion on the total target search time is dependent on the roughness of the energy landscape that the DNA binding protein encounters once it binds non-specifically. We have shown how to determine the escape time for a 3-nt complementary target. This can be extended to estimate the escape time for any complementarity and consequently the diffusion constant on DNA with any base composition [**?**]. Here we have inferred a 15.8 s^{-1} escape rate from the 3-nt CTC guide sequence (**Figure ??**), indicating that if a target strand were to consists only of GA in repeating order, the effective diffusion $D = \frac{1}{2}\frac{dx^2}{dt} = \frac{\text{nt}^2}{2(2\cdot k_{\text{esc}}^{-1})} = \text{nt}^2 k_{\text{esc}} = 15.8\frac{\text{nt}^2}{s}$. Changing the number of base-paring nucleotides as well as the identity of nucleotides in the guide/target could provide insights into how sequence variation would affect the rate of diffusion for other nucleic acid proteins. Since the guide strand only provides the specificity needed for accurate targeting, lateral diffusion could be reliant on the non-specific surface interactions with the protein. We envision that the positive surface charge distribution inside the Ago cleft could orient Ago with the guide towards the negatively charged nucleic acid strand (**Figure ??**), thereby promoting target interrogation while traveling along the target strand. It is unknown whether Ago is able to scan each base during this process or whether it skips over nucleotides. For our triple-target construct, we have observed that 90% of the time the middle target traps Ago. It will be of interest to investigate whether this level of effective target trapping is achieved by a low trapping efficiency offset by repeated passes over the target.

For a longer range target search, we have observed that at distances >100 nt separation, the shuttling time remains well below what would be expected for sliding (**Figure ??**). We show that coiling of the ssDNA (persistence length ∼ 1 nm) may bring distant segments in close proximity, allowing intersegmental jumps over longer distances (beyond 30 nt target separation), and so speeding up lateral diffusion. Interestingly, Ago cannot use intersegmental jumps to cover shorter distances, as implied by the sudden increase in shuttling time when the trap separation goes below 30 nt (**Figure ??**). Experimentally, one could further investigate the nature of intersegmental jumps through a combined tweezer-fluorescence single-molecule assay, where forces strong enough to pull on entropically coiled ssDNA can be applied [**?** **?**]. Furthermore, theoretical modelling and additional experiments are required in order to establish to what extent partitioning the search modes on different length scales will allow nucleic acid guided proteins to optimize the search process [**?** **?** **?**] since the absence of cooperative binding was recentley reported for another Ago system [**?**].

We hypothesize that similar target search strategies may be used by Agos from different families, which are structurally and functionally similar [**?**]. For example, in RNA induced transcriptional silencing (RITS), guide-loaded AGO1 binds to a transcript after which other proteins are recruited for heterochromatin assembly [**?** **?**]. Similarly, in the piRNA pathway PIWI proteins associate with piRNA in germline cells to bind and cleave transposon transcripts in the cytoplasm [**?** **?** **?**] or to nascent RNA in the nucleus in order to induce heterochromatin formation [**?**]. In each of these functions, the reliance on guide-complementary for sequential target search likely necessitates the usage of facilitated diffusion strategies to optimize the search time for proper regulation of cell development or gene stability.

5.4. Methods

5.4.1. Purification of CbAgo

The CbAgo gene was codon harmonized for E.coli Bl21 (DE3) and inserted into a pET-His6 MBP TEV cloning vector (Addgene plasmid # 29656) using ligation independent cloning. The CbAgo protein was expressed in E.coli Bl21(DE3) Rosetta^{TM} 2 (Novagen). Cultures were grown at 37 °C in LB medium containing 50μg ml-1 kanamycin and 34μg ml-1 chloramphenicol till an OD600nm of 0.7 was reached. CbAgo expression was induced by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.1mM. During the expression cells were incubated at 18◦C for 16 hours with continues shaking. Cells were harvested by centrifugation and lysed, through sonication (Bandelin, Sonopuls. 30% power, 1s on/2s off for 5min) in lysis buffer containing 20mM Tris-HCl pH 7.5, 250mM NaCl, 5mM imidazole, supplemented with a EDTA free protease inhibitor cocktail tablet (Roche). The soluble fraction of the lysate was loaded on a nickel column (HisTrap Hp, GE healthcare). The column was extensively washed with wash buffer containing 20mM Tris-HCl pH 7.5, 250mM NaCl and 30mM imidazole. Bound protein was eluted by increasing the concentration of imidazole in the wash buffer to 250mM. The eluted protein was dialysed at 4oC overnight against 20mM HEPES pH 7.5, 250mM KCl, and 1mM dithiothreitol (DTT) in the presence of 1mg TEV protease (expressed and purified according to Tropea et al.63) to cleave of the His6-MBP tag. Next the cleaved protein was diluted in 20mM HEPES pH 7.5 to lower the final salt concentration to 125mM KCl. The diluted protein was applied to a heparin column (HiTrap Heparin HP, GE Healthcare), washed with 20mM HEPES pH 7.5, 125mM KCl and eluted with a linear gradient of 0.125-2M KCl. Next, the eluted protein was loaded onto a size exclusion column (Superdex 200 16/600 column, GE Healthcare) and eluted with 20mM HEPES pH 7.5, 500mM KCl and 1mM DTT. Purified CbAgo protein was diluted in size exclusion buffer to a final concentration of 5uM. Aliquots were flash frozen in liquid nitrogen and stored at -80°C.

5.4.2. Purification of His-tagged Lin28b

The protein was prepared following the protocol of Yeom et al. [**?**].Briefly, recombinant Lin28b was prepared by subcloning cDNA with BamHI and XhoI into pET28-a vector (Novagen). Subsequently, the strain was transformed to E. coli BL21-RIL strain. The expression and purification of recombinant Lin28b was performed according to the manufacturer's protocol.

5.4.3. Single molecule experimental setup

Single molecule FRET experiments were performed with an inverted microscope (IX73, Olympus) with prism-based total internal reflection. Excitation of the donor dye Cy3 is done by illuminating with a 532nm diode laser (Compass 215M/50mW, Coherent). A 60X water immersion objective (UPLSAPO60XW, Olympus) was used for collection of photons from the Cy3 and Cy5 dyes on the surface, after which a 532 nm long pass filter (LDP01-532RU-25, Semrock) blocks the excitation light. A dichroic mirror (635 dcxr, Chroma) separates the fluorescence signal which is then projected onto an EM-CCD camera (iXon Ultra, DU-897U-CS0-#BV, Andor Technology). All experiments were performed at an exposure time of 0.1 s at room temperature (22 ± 0.1 °C)

5.4.4. Fluorescent dye labeling of nucleic acid constructs

All DNA constructs were ordered from ELLA Biotech. Nucleic acid constructs that have an internal amino modification were labeled with fluorescent dyes based on the CSHL protocol 65.1 uL of 1 mM of DNA/RNA dissolved in MilliQ H20 is added to 5 uL labeling buffer of (freshly prepared) sodiumbicarbonate (84 mg/10mL, pH 8.5). 1 uL of 20 mM dye (1 mg in 56 uL DMSO) is added and incubated overnight at 4°C in the dark, followed by washing and ethanol precipitation. Concentration of nucleic acid and labeling efficiency was determined with a Nanodrop spectrophotometer.

5.4.5. Single molecule chamber preparation

Quartz slides were coated with a polyethylene-glycol through the use of amino-silane chemistry. This is followed by assembly of microfluidic chambers with the use of double sided scotchtape. For a detailed protocol, we refer to 66. Further improvement of surface quality occurs through 15 min incubation of T50 and 5% Tween20 67 after which the channel is rinsed with 100 μL T50 buffer. Streptavidin (5 mg/mL) was diluted in T50 to 0.1 mg/mL. 50 μL of this solution is then flowed inside the chamber. This is followed by incubation for 1 min followed by rinsing with approximately 10-fold the volume of the chamber with T50 (10 mM Tris-HCl [pH 8.0], 50 mM NaCl). 100 pM of DNA/RNA target with biotin construct is then flushed in the chamber, followed by 1 min incubation. This is followed subsequently by rinsing with T50. The chamber is subsequently flushed with CbAgo buffer, containing 50 mM Tris-HCl [pH 8.0], 1 mM Trolox, 1 mM MnCl2, 100 mM NaCl. Guide-loading of apo-CbAGO occurs by incubation of the protein (10 nM) with 1 nM guide construct in a buffer containing 50 mM Tris-HCl [pH 8.0], 1 mM Trolox, 1 mM MnCl2, 100 mM NaCl, 0.8% glucose at 37°C for 30 min. Following incubation, glucose oxidase and catalase is added (0.1 mg/mL glucose oxidase) after which the sample is flushed in the microfluidic chamber containing the DNA targets.

5.4.6. Lin28 assay

Immobilization of Lin28b occurred in the following way: 50 μl of streptavidin (0.1 mg/mL) in T50 is flowed inside the chamber and incubated for 1 minute. After this, the chamber is rinsed with approximately 100 μL of T50. 1 μl of Anti-6X His tag® antibody (Biotin) diluted 100-fold in T50 and subsequently flowed inside the chamber. After 5 minutes, the chamber is rinsed with 100 μL of T50. Stock of Lin28b (100 μM) is diluted to 100 nM and incubated with the target DNA (10 nM) and 10 mM MgCl2 for 5 minutes, after which the solution is flushed inside the chamber, followed by incubation of 5 minutes. Lastly, the CbAgo buffer is flushed inside the chamber. Guide-loading of apo-CbAgo occurs in the same way as described above (Single molecule chamber preparation) after which the CbAgo:siDNA complex is also flushed inside the chamber.

5.4.7. QUANTIFICATION AND STATISTICAL ANALYSIS

Fluorescence signals are collected at 0.1-s exposure time unless otherwise specified. For 7-nt target separation, 30-ms exposure time is used. Time traces were subsequently extracted through IDL software using a custom script. Prior to data collection, the location of targets (Cy5 labeled) are found by illuminating the sample with the 637nm laser. Through a mapping file, it subsequently collects the individual intensity hotspots in both the donor and acceptor channel and pairs them up through the mapping file, after which the traces are extracted. During the acquisition of the movie, the green laser is used. Only at the end, the red laser is turned on once more to check for photobleaching of the red dye. Traces containing the fluorescence intensity from the donor and acceptor signal are manually preselected occurs through the use of MATLAB (Mathworks), disregarding artefacts caused by non-specific binding, additional binding to neighboring regions and photobleaching.

5.5. Supplementary Information

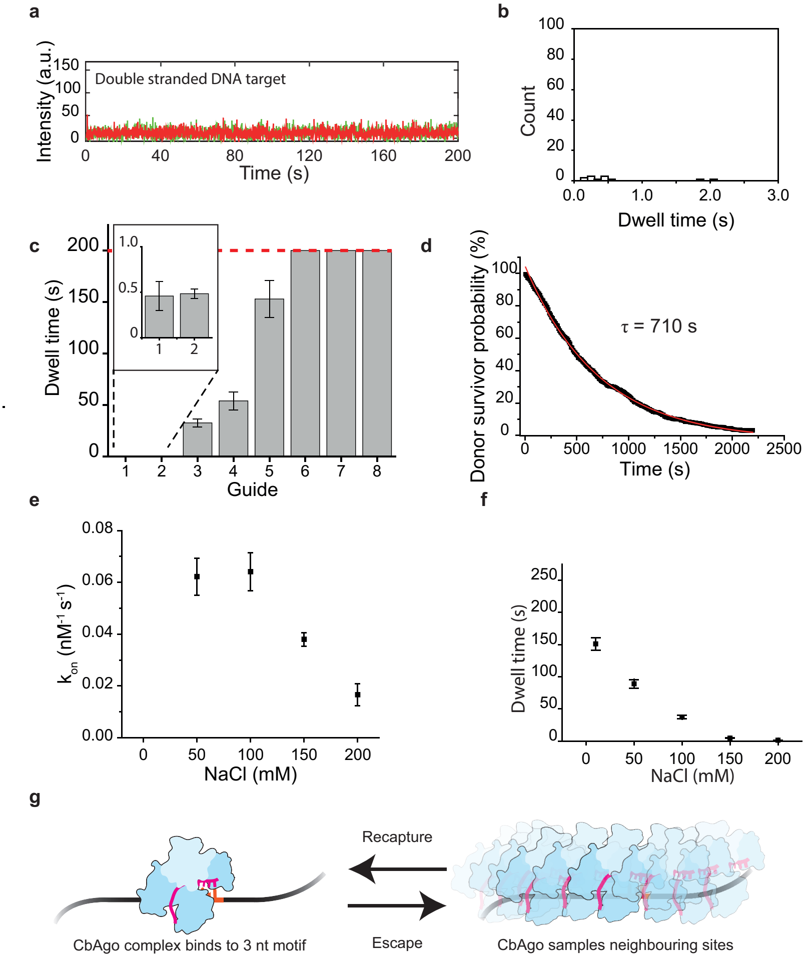


Figure S5.1: Single molecule interactions of CbAgo:siDNA (2-4 nt) at different conditions. (a) Representative trace single-molecule interaction of CbAgo-siDNA (let7) with full target dsDNA target immobilized on the surface (300 per FoV). Exposure time is 100 ms. (b) Dwelltime distribution of CbAgo-guide 3-dsDNA target interactions. Number of molecules recorded n = 540. Number of datapoints n = 12 (c) Average dwell time of protein bound to target versus guide length for N=1 to N=8. The error bars are taken from the 95% confidence interval of bootstrapped dwelltimes (20,000 empirical bootstraps). The striped red line indicates the observation time, limited by photobleaching. (d) Survival plot of donor only (Cy3) constructs in standard experimental conditions (100 mM NaCl, 50 mM Tris-HCl pH 8.0). Mean donor bleaching time was obtained by a single exponential fit to the survival probability plot. (e) Binding rate for different salt concentrations for N=3 (nt 2-4) between guide and single target. (f) Dwell time of CbAgo and a single-stranded single target DNA construct (N=3) at 10, 50, 100, 150 and 200 mM NaCl concentration. Total measurement time = 250 s. Error bars are indicating the 95% percentile of 20,000 empirical bootstraps of the mean dwell time. (G) Schematic image indicating the dynamic escape and recapture events of CbAgo.

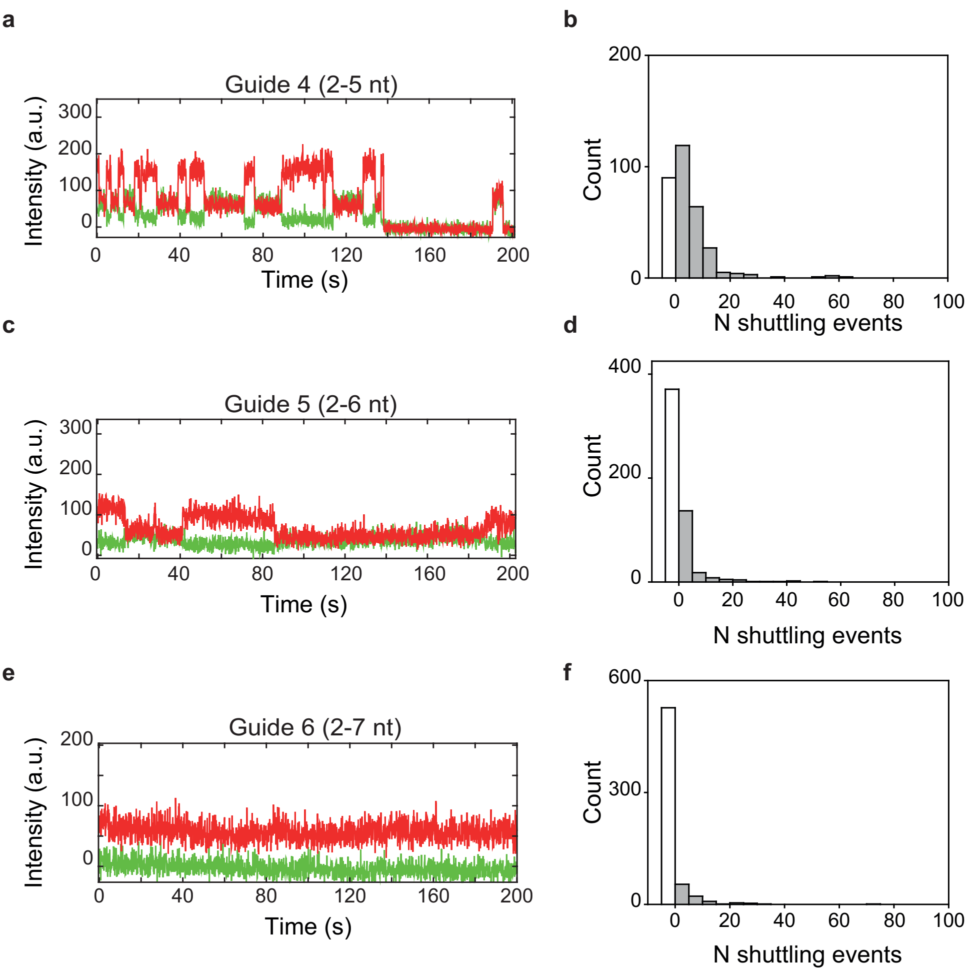


Figure S5.2: Single-molecule interactions of CbAgo with guide 4, 5, 6 and tandem target (22 nt separation). (a) Representative trace of binding events by CbAgo with guide 4 (nt 2-5). Duration of observation 200 s. (b) Shuttling event distribution for guide 4 (nt 2-5). Bin size = 5. The white bar represents binding (no shuttling) events followed by dissociation. N = 317. (c) Representative trace of binding events by CbAgo with guide 5 (2-6). (d) Shuttling event distribution for guide 5 (2-6 nt). Bin size = 10. The white bar represents events that consists of single molecule binding followed by dissociation. n = 550. (e) Representative trace of guide 6 (2-7 nt) interaction.(f) Shuttling event distribution for guide 6. The white bar represents events that consists of single molecule binding followed by dissociation. n = 621.

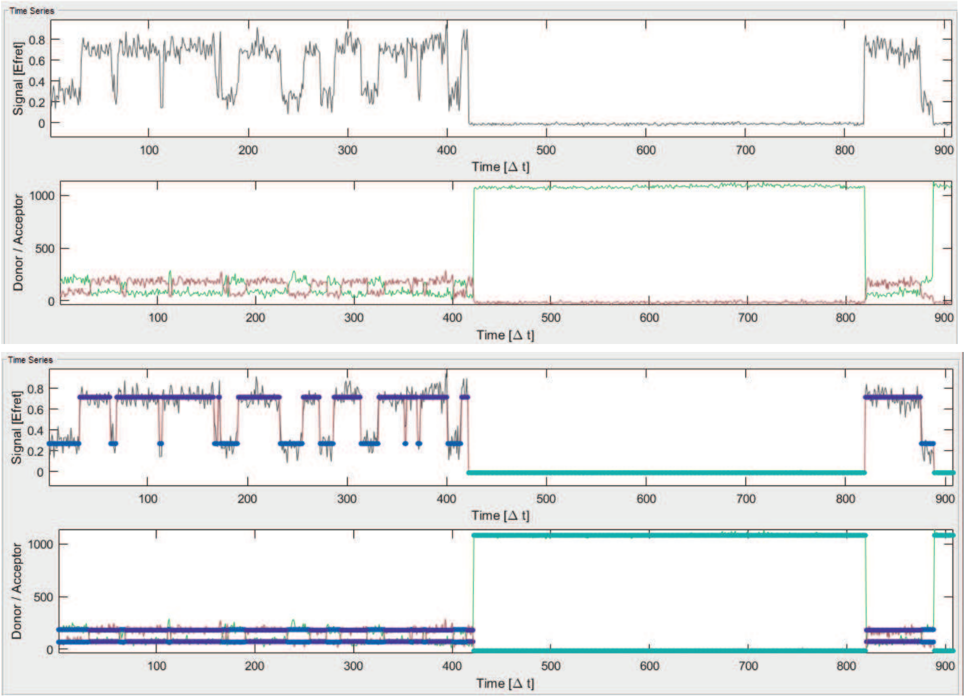


Figure S5.3: Example of HMM software applied to data trace. (Top) An example shuttling trace of CbAgo in the user interface of ebFRET. The donor and acceptor intensities plotted versus time. The donor intensity is enhanced artificially in absence of any signal, resulting in an extra zero FRET state (upper subfigure). (Bottom) The donor, acceptor and FRET intensities overlaid with states resulting from the Hidden Markov Modeling. The HMM analysis program recognizes the unbound state as an extra state (light blue), while low FRET and high FRET are respectively assigned dark blue and purple.

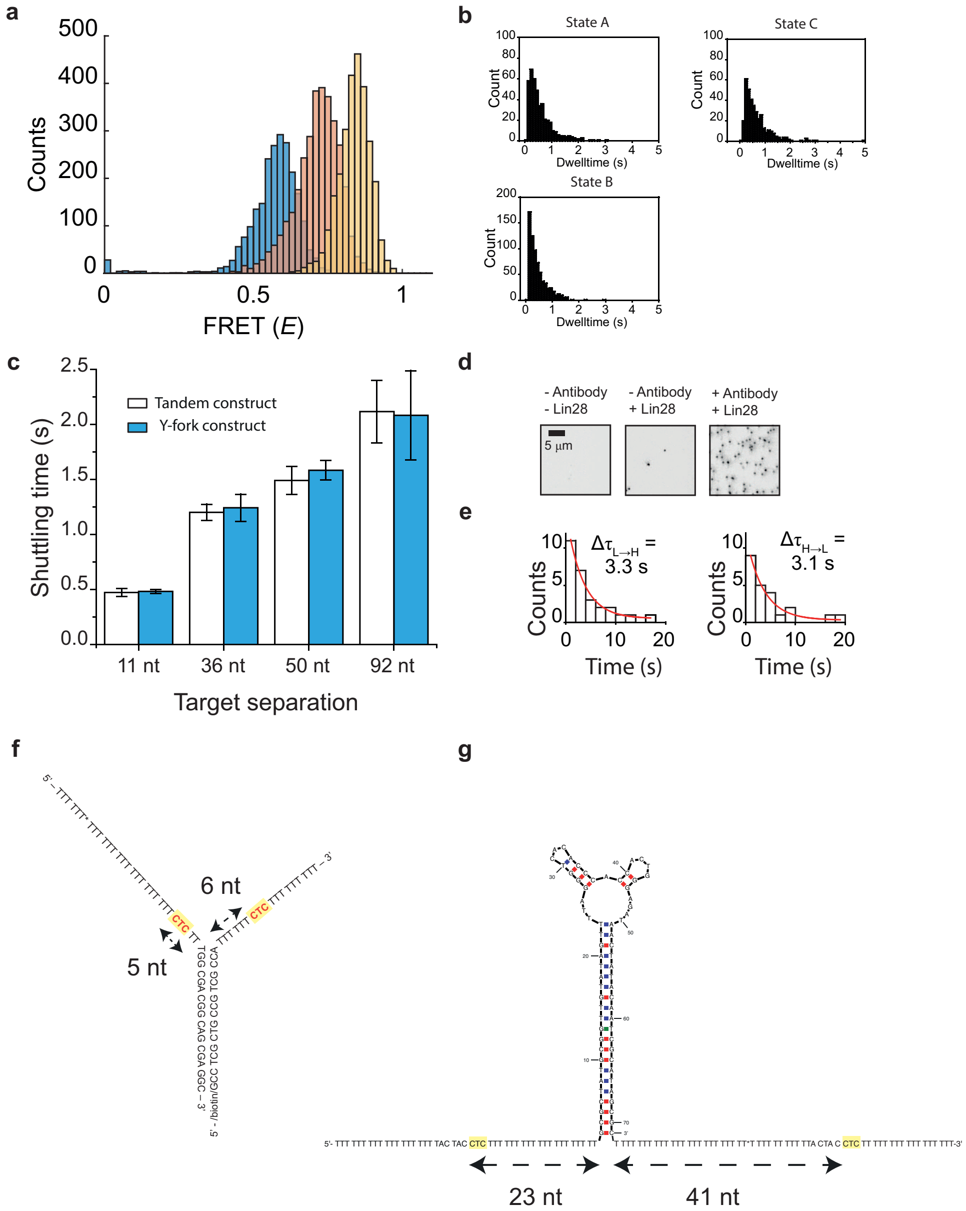


Figure S5.4: Triple target assay, Y-fork assay and Lin28 assay. (a) FRET histogram of three-target assay. n = 168 molecules (b) Dwell time histograms for respectively the low FRET, mid FRET and high FRET state of the three target assay. (c) Shuttling rate of Y-fork constructs (blue) compared to tandem target assay (white) for 11 nt, 36 nt, 50 nt and 92 nt target separation. The error bars indicate the 95% percentile of 20,000 bootstrapped mean dwell times. (d) An EMCCD image of the acceptor channel. (Left) In absence of Lin28 protein and antibody with Cy5 labeled DNA. (Middle) In absence of antibody, but in presence of Lin28 protein and Cy5 labeled DNA. (Right) In presence of antibody, Lin28 protein and Cy5 labeled DNA. (e) Individual dwell times from low FRET state to high FRET state (left) and vice versa (right). (f) Sequence schematic for the Y-fork 11 nt, indicating the target sites and their respective distances to the junction. (g) Sequence schematic for the Lin28 blockade assay, indicating the target sites and their respective distances to the junction/protein.

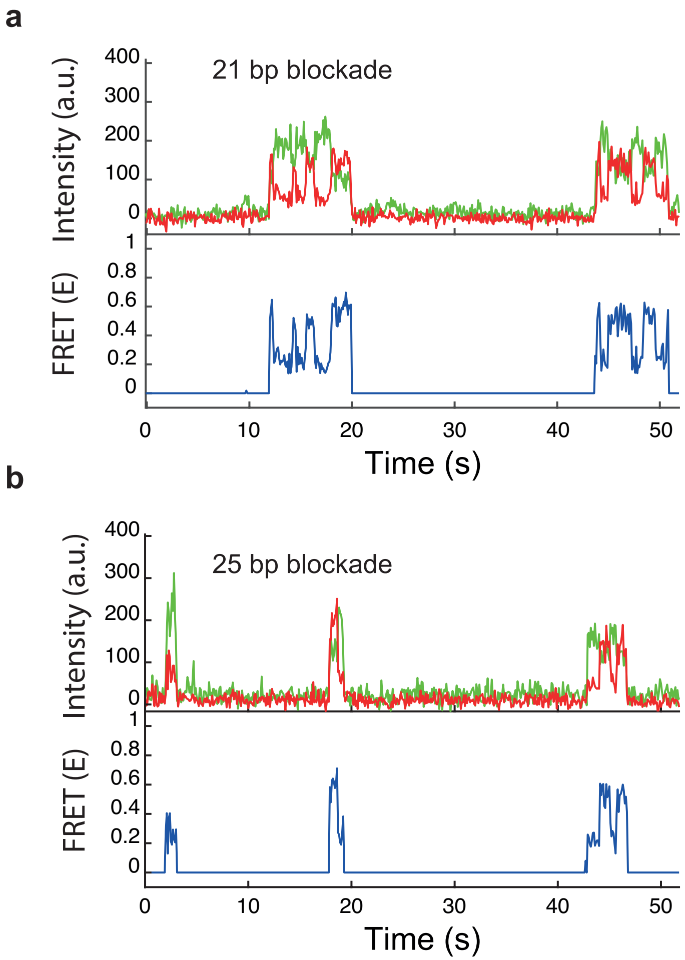


Figure S5.5: Interactions of CbAgo with the dsDNA block construct. (a) Representative trace of CbAgo interacting with a 21 bp DNA blockade construct. (b) Representative trace of CbAgo interacting with a 25 bp DNA blockade construct.

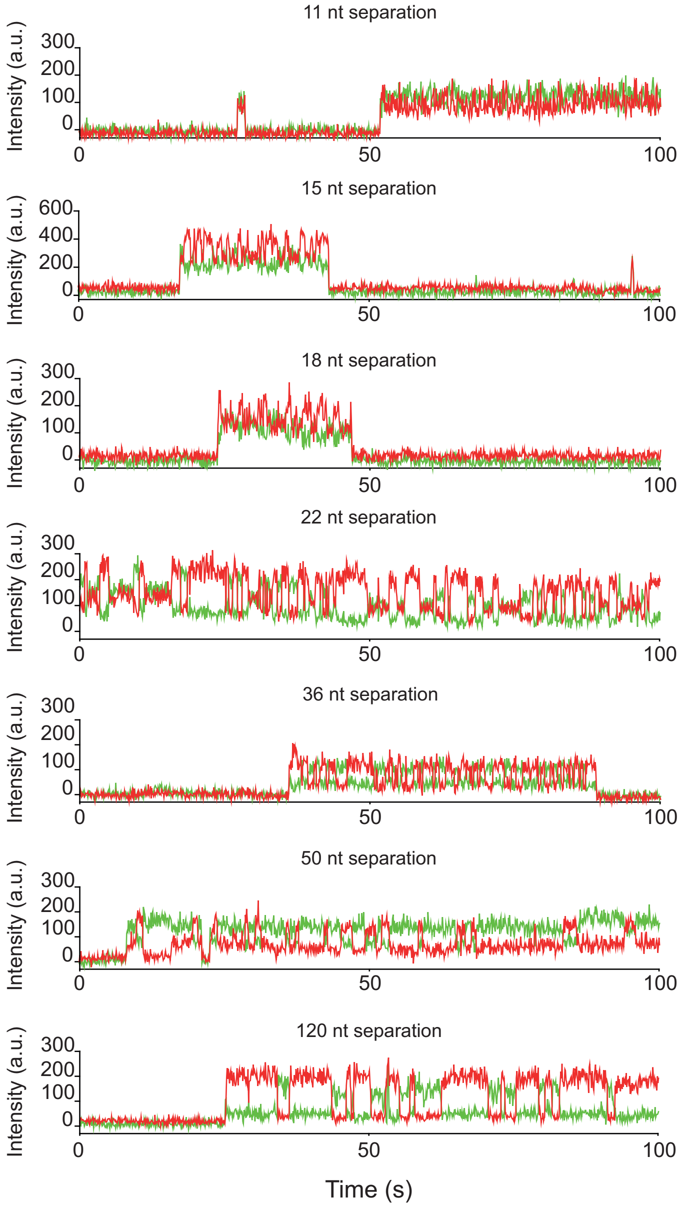


Figure S5.6: Example shuttling traces for 11 nt, 15 nt, 18 nt, 22 nt, 29 nt, 36 nt, 50 nt and 120 nt target separation.

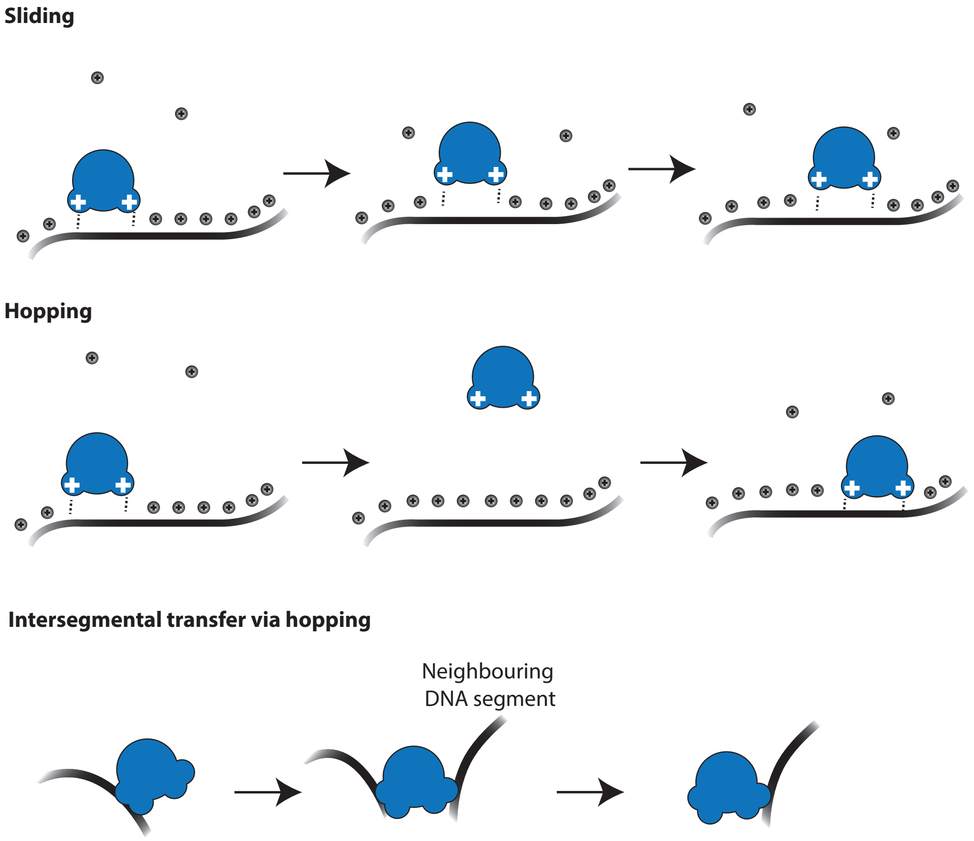


Figure S5.7: Cartoon representation of target search mechanisms. Sliding: Proteins that undergo sliding make a well-correlated movement along the contour of the nucleic acid substrate. There is no net displacement of counterions (grey circles). Hopping: Proteins alternate quickly between a bound and unbound state with respect to DNA and there is counterion condensation upon dissociation of the protein. The method of diffusion is similar to 3D search, but its movements are correlated along the contour of the strand. Intersegmental transfer: This mechanism is a specialized form of hopping where segments appear transiently close by allow the protein to transfer to this new segment.

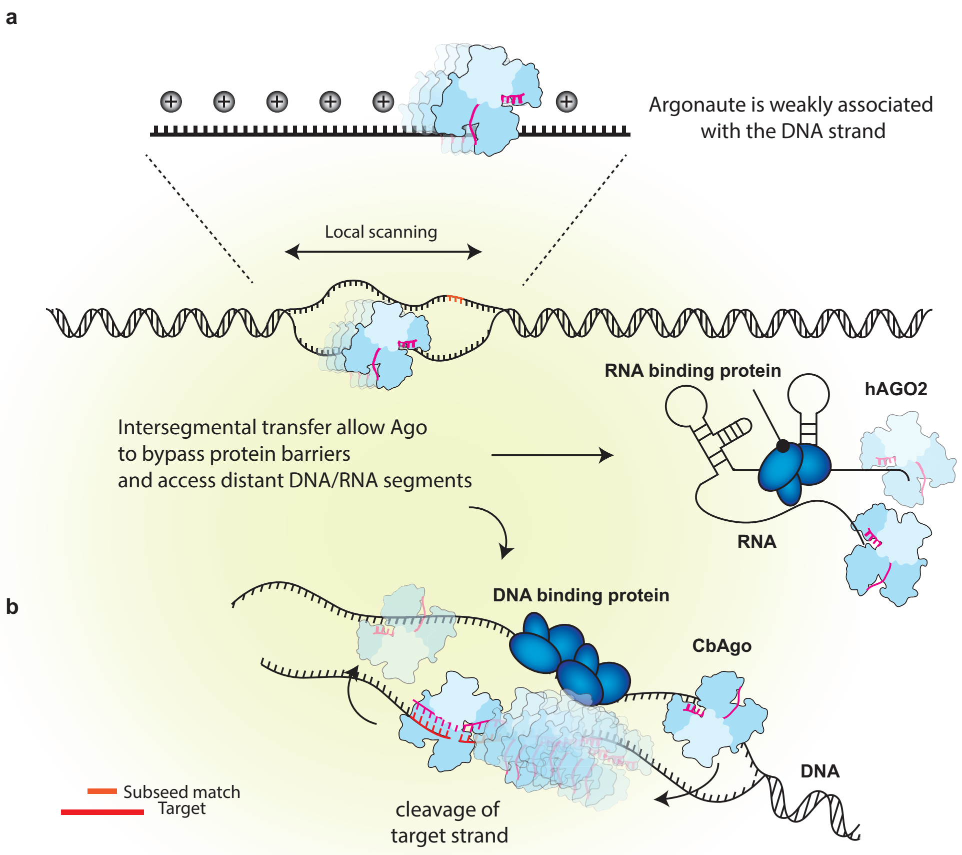

a
Argonaute is weakly associated with the DNA strand
Local scanning
Intersegmental transfer allow Ago to bypass protein barriers and access distant DNA/RNA segments
RNA binding protein
hAGO2
RNA
b
DNA binding protein
CbAgo
DNA
Subseed match
Target
cleavage of target strand

Figure S5.8: Cartoon representation of Ago search model. The Ago complex utilizes short transient interactions with nucleic acid strands to rapidly sample the adjacent (tens of nucleotides away) sites for possible targets. Loose interaction with the nucleic acid strand persists. Obstacles can be overcome through intersegmental jumps.

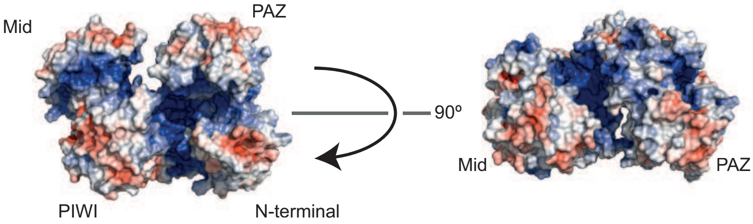


Figure S5.9: Coulombic surface coloring of Clostridium butyricum Argonaute (CbAgo). The crystal structure of CbAgo (PDB 6qzk) (3.23 Å resolution) reveals the charge distribution. The cleft that contains the guide DNA and the target DNA is highly positively charged (blue).

5.5.1. Binding times single-target including recapture events follow single-exponential distribution

We here build a kinetic model for the lateral diffusion by CbAgo. Since Argonaute can in principle bind to any sequence along the DNA, we imagine the binding sites to be located a nucleotide apart. Further, we shall here only explicitly take sliding into account, which is represented as an unbiased random walk with unit step length. Assuming sliding should be a good approximation when considering only short distances traveled. If the protein is bound at the designed 3-nt sub-seed 'target' it can move to either of its neighbors at a rate of k_{esc} or unbind from the ssDNA at a rate of k_{ub}. When bound elsewhere movement and dissociation are assumed to happen instantaneously. To establish the manner in which these undetectable movements contribute to the observed dwell time distribution ($p_{\text{bound}}(\Delta t)$) we count all possible paths that the protein can take to dissociate following initial association to the sub-seed. In Laplace space the unbinding-time distribution, $P_{\text{ub}}(s) = \mathcal{L}\{p_{\text{bound}}(\Delta t)\}$, can be calculated as a product of the distributions of individual transitions (rather than their convolutions), summed over the possible paths towards unbinding. With an exponential distribution of stepping/escape times from the sub-seed trap,

$$p_{\text{esc}}(s) = \frac{2k_{\text{esc}}}{s + 2k_{\text{esc}} + k_{\text{ub}}} \tag{S5.1}$$

, an unbinding time distribution from the trap

$$p_{\text{ub}}(s) = \frac{k_{\text{ub}}}{s + 2k_{\text{esc}} + k_{\text{ub}}} \tag{S5.2}$$

and a probability to return, get recaptured at the trap, from either flank without unbinding P_{retrap} we can write

$$\begin{aligned} P_{\text{ub}}(s) &= \sum_{m=0}^{\infty} \left(p_{\text{esc}}(s)P_{\text{retrap}}\right)^m \left[p_{\text{ub}}(s) + p_{\text{esc}}(s)(1 - P_{\text{retrap}})\right] \\ &= \frac{k_{\text{ub}} + 2k_{\text{esc}}(1 - P_{\text{retrap}})}{s + k_{\text{ub}} + 2k_{\text{esc}}(1 - P_{\text{retrap}})} \end{aligned} \tag{S5.3}$$

The sum on the left hand side of **Equation ??** therefore accounts for the protein escaping from, and getting recaptured at the target an arbitrary amount of times (see **Figure ??** below). The two terms outside the sum represent the probability distributions to unbind from either the target directly or after having escaped one final time respectively (**Figure ??** below). Taking the inverse Laplace transform, we derive the observed dwell time distribution.

$$\begin{aligned} p_{\text{bound}}(\Delta t) &= \mathcal{L}^{-1}\left\{\frac{k_{\text{ub}} + 2k_{\text{esc}}(1 - P_{\text{retrap}})}{s + k_{\text{ub}} + 2k_{\text{esc}}(1 - P_{\text{retrap}})}\right\} \\ &= (k_{\text{ub}} + 2k_{\text{esc}}(1 - P_{\text{retrap}}))e^{-(k_{\text{ub}} + 2k_{\text{esc}}(1 - P_{\text{retrap}}))\Delta t} \end{aligned} \tag{S5.4}$$

Hence, despite the multitude of possible bound states along the DNA the protein can reside in, the observed distribution remains single-exponential. The apparent dissociation rate

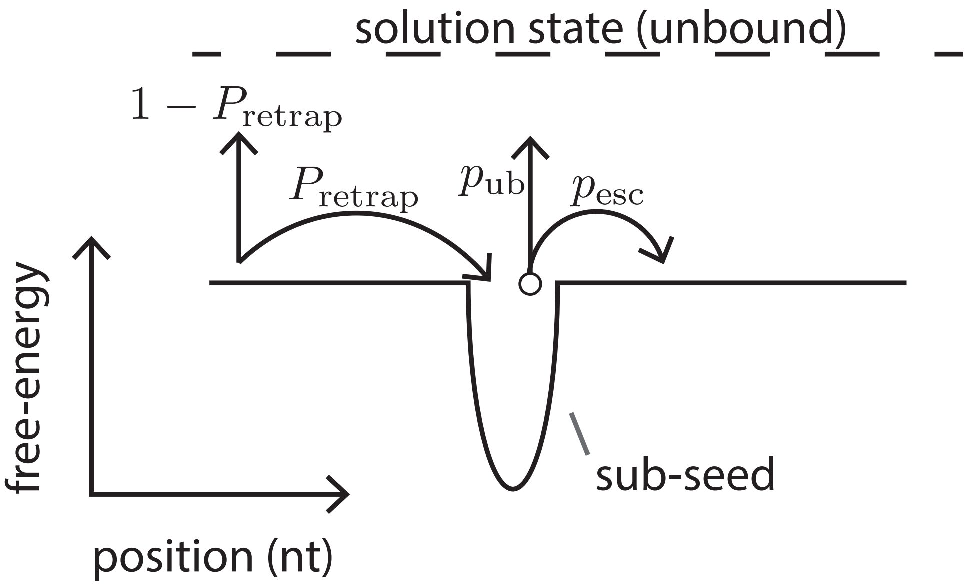


Figure S5.10: This figure illustrates how to construct **Equation ??**. Starting from the sub-seed, Ago can either unbind directly (probability p_{ub}) or slide onto the non-specific binding sites flanking the trap (probability p_{esc}). When non-specifically bound, Ago can either laterally diffuse back into the sub-seed (probability P_{retrap}), or unbind (probability $1 - P_{\text{retrap}}$)

follows

$$k_{\text{ub}}^{\text{observed}} = k_{\text{ub}} + 2k_{\text{esc}}(1 - P_{\text{retrap}}) \tag{S5.5}$$

Given the assay selects for events that get (re-)captured, the observed rate is greater than its intrinsic value.

5.5.2. Shuttling rate due to sliding alone

We seek to explain to what extend sliding contributes to the observed shuttling rate from the tandem-target assay. Given under the current experimental conditions about 13 shuttle events occur prior to unbinding, we shall ignore unbinding in the following analysis ($k_{\text{ub}} \ll k_{\text{esc}}$). To get the distribution of shuttle times ($p(\Delta t_{\text{shuttle}})$) we count all possible paths that lead the protein from one sub-seed to the other. If the two 3-nt nucleotide long sub-seeds are separated by $x_{\text{poly-T}}$ thymine nucleotides, the shuttle times are distributed as (setting $x_{\text{target}} = x_{\text{poly-T}} + 3 \geq 3$) (see **Figure ??** below).

$$\begin{aligned} P_{\text{shuttle}}(s, x_{\text{target}}) &= \sum_{m=0}^{\infty} \left(p_{\text{esc}}(s) \left(\frac{1}{2} \times 1 + \frac{1}{2} \times P_{\text{R}}(x_{\text{target}}) \right) \right)^m p_{\text{esc}}(s) P_{\text{S}}(x_{\text{target}}) \\ &= \frac{k_{\text{esc}} P_{\text{S}}(x_{\text{target}})}{s + k_{\text{esc}} P_{\text{S}}(x_{\text{target}})} \end{aligned} \tag{S5.6}$$

5.5.3. Shuttling rate triple-target construct

For the assay using three sub-seed targets, we can now predict both the time needed to slide from any of the outer ones to the inner ($C \rightarrow B$) and the average time needed to slide along the opposite path ($B \rightarrow C$). The former is equal to the time measured on the tandem target construct, denoted above as $\Delta t_{\mathrm{shuttle}}$ (**Equation ??**, $\Delta\tau_{\mathrm{CB}} = \Delta t_{\mathrm{shuttle}}$). We obtain $\Delta\tau_{\mathrm{BC}}$, via the distribution of lifetimes in the middle trap

$$P(\text{leave } B|\text{arrive at } C)(t) = \frac{P(\text{leave } B)(t)}{P(\text{arrive at } C(\text{and not } A))} \tag{S5.7}$$

Using that the distance between A and B is equal to that in between B and C, in Laplace space, the time spent at target B is distributed as ($P_{\mathrm{B}}(t) \equiv P(\text{leave } B)(t)$)

$$P_{\mathrm{B}}(s, x_{\mathrm{target}}, k_{\mathrm{esc}}) = \sum_{m=0}^{\infty}\left(\frac{1}{2}p_{\mathrm{esc}}(s) \times 2 \times P_{\mathrm{R}}(x_{\mathrm{target}})\right)^{m} \frac{1}{2}p_{\mathrm{esc}}(s)P_{\mathrm{S}}(x_{\mathrm{target}}) \tag{S5.8}$$

The sum accounts for all paths that return to target B. Given the equal distances between all targets on the construct the probability to not make it across to either A or C are equal, which gives rise to the factor of two. The factor outside the sum accounts for the fact that the protein must eventually leave B and make it across to either A or C. Using the same technique as shown above, the average time spent in B equals

$$\tau_{\mathrm{B}}(x_{\mathrm{target}}) = \frac{x_{\mathrm{target}}}{4k_{\mathrm{esc}}} \tag{S5.9}$$

Using that half of the times the protein arrives at A, rather than C, results in the average dwelltime/shuttling time conditioned on moving from B to C (using eq. **??**):

$$\Delta\tau_{\mathrm{BC}}(x_{\mathrm{target}}) = 2\tau_{\mathrm{B}}(x_{\mathrm{target}}) = \frac{x_{\mathrm{target}}}{2k_{\mathrm{esc}}} \tag{S5.10}$$

5.5.4. error estimates using bootstrapping

Fitting the data from the tandem target assay to **Equation ??** provides the estimate of k_{esc}. We bootstrapped the dwell time distributions acquired using the original tandem target assay (distances of 11nt, 15nt, 18nt and 22nt). For each of the 10^5 bootstrap samples we calculated new values for the associated $\Delta t_{\mathrm{shuttle}}$'s and repeated the fit to **Equation ??** to obtain an error estimate in the fitted value of the escape rate.

After using the data from the tandem target assay to estimate k_{esc} there are no more free parameters remaining when predicting the data for the triple-target assay. Performing the bootstrap procedure for k_{esc}, and using **Equations ??** and **??** results in the 95% confidence intervals shown in figure 4D in the main manuscript.

An error estimate for the experimental values of $\Delta\tau_{\mathrm{BC}}$ and $\Delta\tau_{\mathrm{CB}}$ were obtained using 10^5 bootstrap samples of the dwell time distributions measured using the triple-target assay. All analysis was performed with a custom code written in Python. The two terms within the sum shown above represent recapture events at the initial trap via either the the flanking sequence (from which it always returns) or the poly-T stretch in between the traps (from which it returns with a probability $P_{\mathrm{R}}(x_{\mathrm{target}})$ without shuttling) (**Figure ??** shown below).

Finally, the term outside the sum accounts for successful shuttling events (which occurs with probability $P_\mathrm{S}(x_\mathrm{target}) = 1 - P_\mathrm{R}(x_\mathrm{target})$). Once the protein has left the initial trap $P_\mathrm{R}(x)$ and $P_\mathrm{S}(x)$ denote the distributions for either returning back to the initial trap or shuttling/making it across to the other, if the two traps are x nucleotides apart (see **Figure ??** below)). Inverting the Laplace transformation of **Equation ??** we obtain

$$\begin{aligned} p(\Delta t_\mathrm{shuttle}) &= \mathcal{L}^{-1}\left\{\frac{k_\mathrm{esc}P_\mathrm{S}(x_\mathrm{target})}{s + k_\mathrm{esc}P_\mathrm{S}(x_\mathrm{target})}\right\} \\ &= k_\mathrm{esc}P_\mathrm{S}(x_\mathrm{target})e^{-(k_\mathrm{esc}P_\mathrm{S}(x_\mathrm{target})\Delta t_\mathrm{shuttle})} \end{aligned} \tag{S5.11}$$

Hence, the observed dwell time distributions are indeed single exponential. In terms of the microscopic model the average time is set by the escape rate from the trap modified by the probability to make it across once outside of it ($P_\mathrm{S}(x_\mathrm{target})$).
The probabilities P_R and P_S, for a given inter-trap distance x_target follow (see **Figure ??** below)

$$P_\mathrm{R}(x_\mathrm{target}) = \sum_{m=0}^{\infty}\left(\frac{1}{2}P_\mathrm{R}(x_\mathrm{target}-1)\right)^m \frac{1}{2} \tag{S5.12}$$

$$P_\mathrm{S}(x_\mathrm{target}) = \sum_{m=0}^{\infty}\left(\frac{1}{2}P_\mathrm{R}(x_\mathrm{target}-1)\right)^m \frac{1}{2}P_\mathrm{S}(x_\mathrm{target}-1) \tag{S5.13}$$

- from which we can write the recurrence relation

$$P_\mathrm{S}(x_\mathrm{target}) = P_\mathrm{R}(x_\mathrm{target})P_\mathrm{S}(x_\mathrm{target}-1) \tag{S5.14}$$

Using ($P_\mathrm{S}(x_\mathrm{target}) = 1 - P_\mathrm{R}(x_\mathrm{target})$) the above can be re-written as

$$P_\mathrm{S}(x_\mathrm{target}) = \frac{P_\mathrm{S}(x_\mathrm{target}-1)}{P_\mathrm{S}(x_\mathrm{target}-1)+1} \tag{S5.15}$$

which subjected to the boundary condition $P_\mathrm{S}(1) = 1$ - signifying that if the traps are placed adjacent to each other, the shuttle is complete once the protein escaped the initial trap - has the simple solution

$$P_\mathrm{S}(x_\mathrm{target}) = \frac{1}{x_\mathrm{target}} \tag{S5.16}$$

Taken together, the observed shuttling time equals

$$\Delta\tau_\mathrm{shuttle} = \frac{1}{k_\mathrm{esc}P_\mathrm{S}(x_\mathrm{target})} = \frac{x_\mathrm{target}}{k_\mathrm{esc}} \tag{S5.17}$$

Note that $x_\mathrm{target} \geq 3$, as the two sub-seeds cannot overlap. A fit of **Equation ??** to the experimental data for x_target of 11nt, 15nt, 18nt and 22nt in **Figure ??** of the main manuscript were used to estimate the value of k_esc for CbAgo.

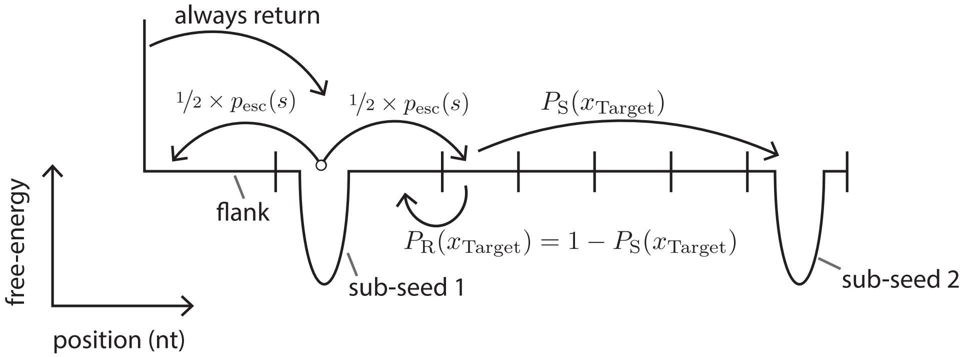

always return
$1/2 \times p_{\text{esc}}(s)$
$1/2 \times p_{\text{esc}}(s)$
$P_{\text{S}}(x_{\text{Target}})$
free-energy
flank
$P_{\text{R}}(x_{\text{Target}}) = 1 - P_{\text{S}}(x_{\text{Target}})$
sub-seed 1
sub-seed 2
position (nt)

Figure S5.11: This figure illustrates how to construct **Equation ??**. Ago slides to either of its neighboring sites with equal probability. Every shuttle event starts with Ago bound to one of the sub-seed sequences. After residing there for a time distributed as $p_{\text{esc}}(s)$, half of the times Ago moves onto the flank (from which it always returns by assumption), while the other half of the times the protein slid onto the poly-T sequence in between the two sub-seeds. All movements along these intermediate sites occur too fast to observe, which is why we only take into account to probability $P_{\text{S}}(x_{\text{target}})$ of completing the shuttle event when x_{target} sites separate Ago from the second sub-seed.

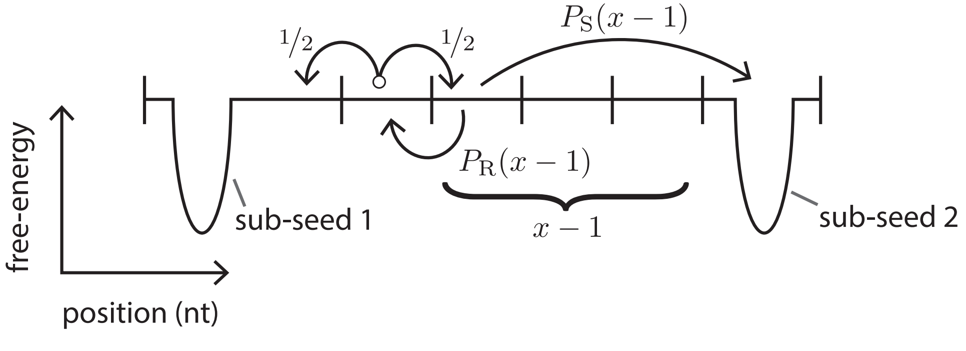


Figure S5.12: This figure illustrates how to construct **Equations ??** and **??**. Let $P_{\text{S}}(x)$ denote the probability to complete the shuttle when x sites separate Ago from the second sub-seed. Ago walks to either of its neighboring sites with equal probability. Therefore, when situated next to the first sub-seed, Ago gets recaptured half of the times it makes a move, while the other half has a probability of $P_{\text{S}}(x-1)$ to result in a completed shuttle event.

5.5.5. Supplementary Tables

Table S1: Dwell times of different two target DNA constructs for several distances. The upper bound and lower bound are estimated through 20000 bootstraps of the acquired dwell times.

Target distance (nt)	Lifetime (sec)	Lower bound lifetime (sec)	Upper bound lifetime (sec)	Shuttling rate (sec-1)	Lower bound shuttling rate (sec-1)	Upper bound shuttling rate (sec-1)
11	0.47	0.46	0.49	2.11	2.04	2.19
15	0.83	0.81	0.87	1.19	1.15	1.24
18	1.17	1.11	1.24	0.85	0.81	0.90
22	1.79	1.74	1.86	0.56	0.54	0.57
29	1.36	1.30	1.42	0.73	0.7	0.77
36	1.19	1.16	1.23	0.84	0.81	0.86
50	1.52	1.46	1.57	0.66	0.64	0.68
64	1.65	1.59	1.71	0.61	0.59	0.63
92	1.94	1.85	2.02	0.52	0.49	0.54
120	2.11	2.03	2.19	0.47	0.46	0.49

Table S2: Oligonucleotides used for this study

Name Oligo	Sequence 5’->3’	Length (nt)
GUIDE		
Guide 3nt (2-4)	5- /5Phos/CGA GTA TT/iAmMC6T/ TTT TTT TTT TTT T – 3’	22
Guide 4nt (2-5)	5’-/5Phos/CGA GGA TT/iAmMC6T/ TTT TTT TTT TTT T - 3’	22
Guide 5nt (2-6)	5’- /5Phos/CGA GGT TT/iAmMC6T/ TTT TTT TTT TTT T - 3’	22
Guide 6nt (2-7)	5’- /5Phos/CGA GGT AT/iAmMC6T/ TTT TTT TTT TTT T - 3 ’	22
Guide 7nt (2-8)	5’- /5Phos/CGA GGT AGA /iAmMC6T/TT TTT TTT TTT T -3’	22
Guide 8nt (2-9)	5’- /5Phos/ CGA GGT AG/iAmMC6T/ TTT TTT TTT TTT T - 3 ’	22
TARGET		
8nt tandem target 7nt separation	5’ - TTT TTT TTT TTT TTT TTT CTC TTT TCT CT/iAmMC6T/ TTT TTT TTT TTT TTT TTT TTT TTT TTT T/biotin/ -3’	58
8nt tandem target 11nt separation	5’ - TTT TTT TTT TTT TTT TTT CTC TTT TTT TT CT CT/iAmMC6T/ TTT TTT TTT TTT TTT TTT TTT TTT TTT T/biotin/ -3’	62
8nt tandem target 15nt separation	5’ - TTT TTT TTT TTT TAC TAC CTC TTT TTT TA CTA CCT CT/iAmMC6T/ TTT TTT TTT TTT TTT TTT TTT TTT TTT T/biotin/ -3’	66
8nt tandem target 18nt separation	5’ - TTT TTT TTT TTT TAC TAC CTC TTT TTT TTT TA CTA CCT CT/iAmMC6T/ TTT TTT TTT TTT TTT TTT TTT TTT TTT T/biotin/ -3’	69
8nt tandem target 22nt separation	5’ - TTT TTT TTT TTT TAC TAC CTC TTT TTT /iAmMC6T/TT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT T/biotin/ -3’	73
8nt tandem target 29nt separation	5’ –TTT TTT TTT TTT TA CTA CCT CTT TT TTT TT/iAmMC6T/ TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/biotin/-3’	81
8nt double target 36nt separation	5’ –TTT TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TT/iAmMC6T/ TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/biotin dT//Phos/-3’	89

8 nt tandem target 50nt separation	5’ –TTT TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TT TTT TTT TT/iAmMC6T/ TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/biotin dT//Phos/-3’	104
8 nt tandem target 64 nt separation	5’ –TTT TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT T/iAmMC6T/T TTT TTT TTT TTT ACT ACC TCT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/biotin-dT/ /Phos/-3’	117
8 nt tandem target 92 nt separation	5’ –TTT TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT T TTT TTT TTT TTT TTT TT TTT TTT TTT TTT TTT TTT TTT T/iAmMC6T/T TTT TTT TTT T ACT ACC TCT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/biotin-dT/ /Phos/-3’	145
8nt double target 120nt separation	5’ –TTT TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/iAmMC6T/ TTT TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/biotin dT//Phos/-3’	171
11nt Y-fork	5’ – TTT TTT* TTT TTT TTT TTT TTT TTT CTC TT TGG CGA CGG CAG CGA GGC – 3’	47
11nt Y-fork biotin	5’ - /biotin/GCC TCG CTG CCG TCG CCA TTT TTT CTC TTT TTT TTT – 3’	36
50nt Y-fork	5‘- TTT TTT TTT TTT* TTT TTT TAC TAC CTC TTT TTT TTT TTT TTT TT TTT TGG CGA CGG CAG CGA GGC – 3‘	65
Y-fork stem (not for Y11)	5’ – /biotin/GCC TCG CTG CCG TCG CCA TTT TTT TTT TTT TTT TTT TTT TAC TAC CTC TTT TTT TTT – 3’	57
36nt dsDNA target	5’ – TTT TTT TTT TTT T TA CTA C CTC T CGG ACC AAC AGC GGG /T-biotin/AC GGC TGT GC TA CTA CCT CTT TTT TTT TTT TTT TTT TTT - 3’	78
36nt dsDNA block v2 3’ biotin	5’ –TTT TTT TTT TTT T TA CTA C CTC T CGG ACC AAC AGC GGG TAC GGC TGT GC TA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/biotin dT/- 3’	91
36nt dsDNA block 5’end truncated	5’ –CTA C CTC T CGG ACC AAC AGC GGG TAC GGC TGT GC TA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT/biotin dT/- 3’	75
36nt 25nt block	5’ - TA GC ACA GCC GT* A CCC GCT GTT GGT- 3’	25
36nt 21nt block	5’- GC ACA GCC GT* A CCC GCT GTT G- 3’	21
36nt 17nt block	5’ -ACA GCC GT* A CCC GCT GT- 3’	17

Triple target	5’ – T/iAmMC6T/ TTT TTT TTT TAC CTC TTT TTT ACC TCT TTT TTA CCT C TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT/biotin/ -3’	69
No target DNA	5’- TTT TTT TTT TTT TTT TTT TTT TTT TTT /iAmMC6T/TT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGG CGA CGG CAG CGA GGC -3’	90
8nt single target	5’ - TTT TTT TTT TTT TTT TTT TTT TTT TTT /iAmMC6T/TT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT T/biotin/-3’	73
3’ biotin stem	5’ - GCC TCG CTG CCG TCG CCA biotin – 3’	18
Lin28 double target	5’- TTT TTT TTT TTT TTT TTT TAC TAC CTC TTT TTT TTT TTT TTT TTT TTG CGC TAT GCG GTT GTA TAG TTT TAG GGT CAC ACC CAC CAC TGG GAG ATA ACT ATA CAA TCG CAT AGC GCT TTT TTT TTT TTT TTT TTT TTT T/iAmMC6T/T TTT TTT TTA CTA CCT CTT TTT TTT TTT TTT TTT-3’	174

References

[] A. Eulalio, E. Huntzinger, and E. Izaurralde, *Getting to the Root of miRNA-Mediated Gene Silencing,* Cell **132**, 9 (2008).

[] D. C. Swarts, K. Makarova, Y. Wang, K. Nakanishi, R. F. Ketting, E. V. Koonin, D. J. Patel, and J. van der Oost, *The evolutionary journey of Argonaute proteins.* Nature structural & molecular biology **21**, 743 (2014).

[] L. He and G. J. Hannon, *MicroRNAs: Small RNAs with a big role in gene regulation,* Nature Reviews Genetics **5**, 522 (2004).

[] F. V. Rivas, N. H. Tolia, J.-J. Song, J. P. Aragon, J. Liu, G. J. Hannon, and L. Joshua-Tor, *Purified Argonaute2 and an siRNA form recombinant human RISC,* Nature structural & molecular biology **12**, 340 (2005).

[] D. P. Bartel, *MicroRNAs: Target Recognition and Regulatory Functions,* Cell **136**, 215 (2009), arXiv:0208024 [gr-qc] .

[] I. Olovnikov, K. Chan, R. Sachidanandam, D. K. D. Newman, and A. A. A. Aravin, *Bacterial Argonaute Samples the Transcriptome to Identify Foreign DNA,* Molecular Cell **51**, 594 (2013), arXiv:NIHMS150003 .

[] D. C. Swarts, M. M. Jore, E. R. Westra, Y. Zhu, J. H. Janssen, A. P. Snijders, Y. Wang, D. J. Patel, J. Berenguer, S. J. J. Brouns, and J. van der Oost, *DNA-guided DNA interference by a prokaryotic Argonaute.* Nature **507**, 258 (2014), arXiv:15334406 .

[] D. C. Swarts, J. van der Oost, and M. Jinek, *Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a,* Molecular Cell **66**, 221 (2017).

[] L. B. Harrington, L. B. Harrington, D. Burstein, J. S. Chen, D. Paez-espino, E. Ma, I. P. Witte, J. C. Cofsky, N. C. Kyrpides, J. F. Banfield, and J. A. Doudna, *Programmed DNA destruction by miniature CRISPR-Cas14 enzymes,* Science **4294**, 1 (2018).

[] R. C. Friedman, K. K. H. Farh, C. B. Burge, and D. P. Bartel, *Most mammalian mRNAs are conserved targets of microRNAs,* Genome Research **19**, 92 (2009).

[] D. Kim, Y. M. Sung, J. Park, S. Kim, J. Kim, J. Park, H. Ha, J. Y. Bae, S. Kim, and D. Baek, *General rules for functional microRNA targeting,* Nature Genetics **48**, 1517 (2016).

[] O. G. Berg, R. B. Winter, and P. H. von Hippel, *Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory.* Biochemistry **20**, 6929 (1981).

[] A. D. Riggs, H. Suzuki, and S. Bourgeois, *lac repressor-operator interaction: I. Equilibrium studies,* Journal of Molecular Biology **48**, 67 (1970).

[] S. D. Chandradoss, N. T. Schirle, M. Szczepaniak, I. J. Macrae, and C. Joo, *A Dynamic Search Process Underlies MicroRNA Targeting,* Cell **162**, 96 (2015).

[] U. Gerland, J. D. Moroz, and T. Hwa, *Physical constraints and functional characteristics of transcription factor-DNA interaction,* Proceedings of the National Academy of Sciences **99**, 12015 (2002), arXiv:0112083 [physics] .

[] A. B. Kolomeisky and A. Veksler, *How to accelerate protein search on DNA: Location and dissociation,* Journal of Chemical Physics **136** (2012), 10.1063/1.3697763.

[] B. van den Broek, M. A. Lomholt, S.-M. J. Kalisch, R. Metzler, and G. J. L. Wuite, *How DNA coiling enhances target localization by proteins.* Proceedings of the National Academy of Sciences of the United States of America **105**, 15738 (2008).

[] S. E. Halford and J. F. Marko, *How do site-specific DNA-binding proteins find their targets?* Nucleic Acids Research **32**, 3040 (2004).

[] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf, *The lac Repressor Displays Facilitated Diffusion in Living Cells,* Science **336**, 1595 (2012).

[] P. H. Von Hippel and O. G. Berg, *Facilitated target location in biological systems,* Journal of Biological Chemistry **264**, 675 (1989).

[] V. Globyte, S. H. Kim, and C. Joo, *Single-Molecule View of Small RNA–Guided Target Search and Recognition,* Annual Review of Biophysics **47**, 569 (2018).

[] D. Singh, S. H. Sternberg, J. Fei, J. A. Doudna, and T. Ha, *Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9,* Nature Communications **7**, 1 (2016).

[] C. Xue, Y. Zhu, X. Zhang, Y. K. Shin, and D. G. Sashital, *Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade,* Cell Reports **21**, 3717 (2017).

[] D. L. Jones, P. Leroy, C. Unoson, D. Fange, V. Ćurić, M. J. Lawson, and J. Elf, *Kinetics of dCas9 target search in Escherichia coli,* Science **357**, 1420 (2017).

[] C. Li, V. V. Vagin, S. Lee, J. Xu, S. Ma, H. Xi, H. Seitz, M. D. Horwich, M. Syrzycka, B. M. Honda, E. L. Kittler, M. L. Zapp, C. Klattenhoff, N. Schulz, W. E. Theurkauf, Z. Weng, and P. D. Zamore, *Collapse of Germline piRNAs in the Absence of Argonaute3 Reveals Somatic piRNAs in Flies,* Cell **137**, 509 (2009).

[] S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna, *DNA interrogation by the CRISPR RNA-guided endonuclease Cas9,* Nature **507**, 62 (2014), arXiv:NIHMS150003 .

[] V. Globyte, S. H. Lee, T. Bae, J.-S. Kim, and C. Joo, *CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion,* The EMBO Journal , e99466 (2018).

[] W. E. Salomon, S. M. Jolly, M. J. Moore, P. D. Zamore, and V. Serebrov, *Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides,* Cell **162**, 84 (2015).

[] C. S. Smith, K. Jouravleva, M. Huisman, S. M. Jolly, P. D. Zamore, and D. Grunwald, *An automated Bayesian pipeline for rapid analysis of single-molecule binding data,* Nature Communications **10**, 1 (2019).

[] J. W. Hegge, D. C. Swarts, S. D. Chandradoss, T. J. Cui, J. Kneppers, M. Jinek, C. Joo, and J. van der Oost, *DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute,* Nucleic Acids Research **47**, 5809 (2019), arXiv:/doi.org/10.1101/534206 [https:] .

[] K. Ragunathan, C. Liu, and T. Ha, *RecA filament sliding on DNA facilitates homology search,* eLife **2012**, 1 (2012).

[] J. W. Van De Meent, J. E. Bronson, C. H. Wiggins, and R. L. Gonzalez, *Empirical bayes methods enable advanced population-level analyses of single-molecule FRET experiments,* Biophysical Journal **106**, 1327 (2014).

[] B. M. Muhire, M. Golden, B. Murrell, P. Lefeuvre, J.-M. Lett, A. Gray, A. Y. F. Poon, N. K. Ngandu, Y. Semegni, E. P. Tanov, A. L. Monjane, G. W. Harkins, A. Varsani, D. N. Shepherd, and D. P. Martin, *Evidence of Pervasive Biologically Functional Secondary Structures within the Genomes of Eukaryotic Single-Stranded DNA Viruses,* Journal of Virology **88**, 1972 (2014).

[] E. J. Strobel, A. M. Yu, and J. B. Lucks, *High-throughput determination of RNA structures,* Nature Reviews Genetics **19**, 615 (2018).

[] Y. Nam, C. Chen, R. I. Gregory, J. J. Chou, and P. Sliz, *Molecular basis for interaction of let-7 MicroRNAs with Lin28,* Cell **147**, 1080 (2011), arXiv:NIHMS150003 .

[] H. Chen, S. P. Meisburger, S. a. Pabit, J. L. Sutton, W. W. Webb, and L. Pollack, *Ionic strength-dependent persistence lengths of single-stranded RNA and DNA,* Proceedings of the National Academy of Sciences **109**, 799 (2012).

[] M. A. Lomholt, B. van den Broek, S.-M. J. Kalisch, G. J. L. Wuite, and R. Metzler, *Facilitated diffusion with DNA coiling,* Proceedings of the National Academy of Sciences **106**, 8204 (2009).

[] D. Banerjee and F. Slack, *Control of developmental timing by small temporal rnas: A paradigm for rna-mediated regulation of gene expression,* BioEssays **24**, 119 (2002).

[] R. B. Winter and P. H. Von Hippel, *Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli lac repressor-operator interaction: equilibrium measurements,* Biochemistry **20**, 6948 (1981).

[] R. B. Winter, O. G. Berg, and P. H. von Hippel, *Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor–operator interaction: kinetic measurements and conclusions.* Biochemistry **20**, 6961 (1981).

[] C. Desnues, B. Rodriguez-Brito, S. Rayhawk, S. Kelley, T. Tran, M. Haynes, H. Liu, M. Furlan, L. Wegley, B. Chau, Y. Ruan, D. Hall, F. E. Angly, R. A. Edwards, L. Li, R. V. Thurber, R. P. Reid, J. Siefert, V. Souza, D. L. Valentine, B. K. Swan, M. Breitbart, and F. Rohwer, *Biodiversity and biogeography of phages in modern stromatolites and thrombolites,* Nature **452**, 340 (2008).

[] J. M. Labonté and C. A. Suttle, *Previously unknown and highly divergent ssDNA viruses populate the oceans,* ISME Journal **7**, 2169 (2013).

[] M. Yoshida, T. Mochizuki, S. I. Urayama, Y. Yoshida-Takashima, S. Nishi, M. Hirai, H. Nomaki, Y. Takaki, T. Nunoura, and K. Takai, *Quantitative viral community DNA analysis reveals the dominance of single-stranded DNA viruses in offshore upper bathyal sediment from Tohoku, Japan,* Frontiers in Microbiology **9**, 1 (2018).

[] O. Barabas, D. R. Ronning, C. Guynet, A. B. Hickman, B. Ton-Hoang, M. Chandler, and F. Dyda, *Mechanism of IS200/IS605 Family DNA Transposases: Activation and Transposon-Directed Target Site Selection,* Cell **132**, 208 (2008).

[] M. J. Curcio and K. M. Derbyshire, *The outs and ins of transposition: From MU to kangaroo,* Nature Reviews Molecular Cell Biology **4**, 865 (2003).

[] B. Ton-Hoang, C. Pasternak, P. Siguier, C. Guynet, A. B. Hickman, F. Dyda, S. Sommer, and M. Chandler, *Single-stranded DNA transposition is coupled to host replication,* Cell **142**, 398 (2010).

[] J.-D. Beaudoin, E. M. Novoa, C. E. Vejnar, V. Yartseva, C. M. Takacs, M. Kellis, and A. J. Giraldez, *Analyses of mRNA structure dynamics identify embryonic gene regulatory programs,* Nature Structural & Molecular Biology , 1 (2018).

[] P. C. Bevilacqua, L. E. Ritchey, Z. Su, and S. M. Assmann, *Genome-Wide Analysis of RNA Secondary Structure,* Annual Review of Genetics **50**, 235 (2016).

[] Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, and S. M. Assmann, *In vivo genome-wide profiling of RNA secondary structure re-*

veals novel regulatory features, Nature **505**, 696 (2014).

[] L. E. Vandivier, S. J. Anderson, S. W. Foley, and B. D. Gregory, *The Conservation and Function of RNA Secondary Structure in Plants,* Annual Review of Plant Biology **67**, 463 (2016), arXiv:1510.01420 .

[] R. Zwanzig, *Diffusion in a rough potential.* Proceedings of the National Academy of Sciences **85**, 2029 (1988).

[] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj, *How a protein searches for its site on DNA: the mechanism of facilitated diffusion,* Journal of Physics A: Mathematical and Theoretical **42**, 434013 (2009).

[] M. Sheinman and Y. Kafri, *The effects of intersegmental transfers on target location by proteins.* Physical biology **6**, 016003 (2009), arXiv:0807.3639 .

[] M. Slutsky and L. A. Mirny, *Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential,* Biophysical Journal **87**, 4021 (2004), arXiv:0402005 [q-bio] .

[] D. Holoch and D. Moazed, *Small-RNA loading licenses Argonaute for assembly into a transcriptional silencing complex.* Nature structural & molecular biology **22**, 328 (2015).

[] D. Moazed, *Small RNAs in transcriptional gene silencing and genome defence,* Nature **457**, 413 (2009), arXiv:NIHMS150003 .

[] J. Brennecke, A. A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, and G. J. Hannon, *Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila,* Cell **128**, 1089 (2007).

[] L. S. Gunawardane, K. Saito, K. M. Nishida, K. Miyoshi, Y. Kawamura, T. Nagami, H. Siomi, and M. C. Siomi, *A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila,* Science **315**, 1587 (2007).

[] G.-W. Li, O. G. Berg, and J. Elf, *Effects of macromolecular crowding and DNA looping on gene regulation kinetics,* Nature Physics **5**, 294 (2009).

[] A. A. Aravin, R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, K. F. Toth, T. Bestor, and G. J. Hannon, *A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice,* Molecular Cell **31**, 785 (2008), arXiv:NIHMS150003 .

[] K. H. Yeom, I. Heo, J. Lee, S. Hohng, V. N. Kim, and C. Joo, *Single-molecule approach to immunoprecipitated protein complexes: Insights into miRNA uridylation,* EMBO Reports **12**, 690 (2011).

6

Optimal DNA/RNA target search using frequent skip-n-slides

The timed action of target searching proteins at specific DNA or RNA sequences plays a vital role in the cell. A special class of such target searchers, amongst which Argonaute and CRISPR-Cas9, use small RNA or DNA guides to define their target site. These guides can readily be synthesized, enabling the repurposing of the target searching proteins for genome engineering. Here we employ a combination of single-molecule FRET and theoretical modeling to understand the microscopic kinetics underlying the target search. We show both a prokaryotic and an eukaryotic Argonaute only sparsely interrogate their ssDNA/mRNA substrates, using a mixture of sliding to neighboring sites and frequent skipping to interrogate distant sites. Next, we show such a mixture minimizes the time needed to locate the target. Hence, we suggest Argonaute seems to operate at near optimal conditions using a mechanism likely applicable to other (guided and non-guided) target searchers.

manuscript in preparation

6.1. Introduction

A multitude of cellular processes, including gene regulation, DNA repair, and immune response rely on proteins binding to specific DNA or RNA sequences. Even if the protein interacts only with the correct target sequence, the sheer size of the intracellular volume restricts the rate at which it can be found through diffusive collisions alone [**? ? ? ?**]. Still, measured search speeds can exceed the upper limit for diffusive collisions with up to two orders of magnitude [**?**]. To reach the observed speeds, target searching proteins can reduce the effective size of their search space by spending some fraction of time non-specifically associated and diffusively sliding along the DNA—partially replacing excursions into the solution (3D motion) as a means of reaching new sites to interrogate [**? ? ? ?**]. Theoretical work [**?**] showed that an equal split of time spent sliding along the DNA and diffusing through solution would minimize the search time. While experiments have indeed confirmed such facilitated diffusion (a mix of 1D and 3D motion) for a variety of proteins [**? ? ? ? ? ? ? ? ? ? ? ? ? ? ?**], in vivo studies suggested the system spends considerably more than half the time associated to DNA [**? ?**].

Repeated transfer between 1D sliding motion and 3D diffusion through solution and rebinding at an uncorrelated site is beneficial, as the sliding motion by itself will inevitably double back on itself and wastes time interrogating sites already visited. Early theoretical work recognized that this scanning redundancy could be further reduced if the non-specific interactions allow for intersegmental transfers [**? ? ? ?**], where the protein quickly moves between close by DNA segments without fully returning to the solution state [**? ? ?**]. If the search process is optimized for time, and the total time spent transferring between segments is assumed negligible, we expect intersegmental transfers to minimize search redundancy (and so search time) by occurring as frequently as the geometry of the substrate allows. It has been shown theoretically that allowing for a (small) fixed amount of rapid intersegmental transfers shifts the optimal partitioning between 1D and 3D diffusion toward spending more time associated with the DNA [**?**].

However, when such transfers occur frequently the total time spent transitioning between segments cannot be neglected. For instance, we expect this to be the case for proteins searching along single-stranded (ss) RNA or DNA with persistence lengths on the order of one nucleotide (nt) [**?**]. We may expect similar behavior for proteins that bind genomic targets, due to the strongly compacted double-stranded (ds) DNA within the nucleus or bacterial nucleoid. Furthermore, cellular RNA or DNA is typically occupied by various other (non-)specific binding proteins [**? ?**], or can form secondary structures (i.e. plectonemes on dsDNA, or hairpins on ssRNA), all forming roadblocks along the target searcher's path. Bypassing such obstacles is often impossible through sliding, thereby necessitating the frequent use of some form of base-skipping, such as intersegmental transfers in case of sufficiently flexible substrates. Irrespective of the particular mechanism used, bases along the substrate are not interrogated, and we will simply refer to this process as 'skipping'. Little is known of the effect the frequent skips have on the search time.

Here we use Argonaute (Ago) as a model system for searches along flexible ss substrates. Ago belongs to a particular class of target searchers that pair with a small non-coding RNA (or DNA) guide, and then targets its complementary sequence [**?**]. The common usage of the CRISPR-Cas9 and CRISPR-Cas12a as next-generation genome editing tools [**?**], further highlights the importance of understanding also how such guided target searchers operate [**? ? ? ? ? ?**]. A recent study showed the prokaryotic Clostridium butyricum Argonaute (CbAgo) uses a ssDNA guide to cleave ssDNA or dsDNA at moderate temperatures (~ 37 °C) [**?**], making it a suitable candidate as a genome-editing tool. In a previous study [**?**] (**Chapter ??**) we demonstrated CbAgo can bypass roadblocks while diffusing along its substrate. Here, we start by establishing the generality of this base-skipping behavior by confirming its existence also for the eukaryotic human Argonaute 2 (hAgo2), using single-molecule (sm) Förster resonance energy transfer (FRET).

Next, we ask under what conditions skips can speed-up the target search process. To this end, we draw inspiration from established models [**? ? ? ?**] and consider the search as consisting of three parts, but crucially allow all parts to take a finite time to complete: (i) interrogation of off-targets through sliding, (ii) base skipping, and (iii) diffusion through solution, followed by rebinding at an uncorrelated site. Through our modeling we discover the existence of two optimal partitioning between the three search modes: one coinciding with the known optimum of an equal time-split between 1D and 3D diffusion through solution when no skipping is allowed [**?**], and one novel optimum where skipping and sliding coexist during lateral diffusion. We fully characterize the search optima, and show that as a general rule, the system can never spend more time in solution than on the substrate when optimized, in accordance with experimental results [**? ?**].

Using the presented smFRET data, we conclude by arguing that Ago operates far from the sliding-only optimum, and that its search characteristics are consistent with the skip-and-slide optimum. Our work suggests that any search involving many skips soon becomes beneficial over using only sliding, and thus raises the question whether skip-and-slide search could also be the preferred search mode for other searchers.

6.2. Results

6.2.1. Single-molecule FRET assay to probe lateral diffusion

Diffusive motion is often characterized by measuring the mean square displacement as a function of time [**? ? ? ? ? ? ? ?**]. Even in the best of scenarios, when considering a stretched and uncoiled substrate, direct observation of lateral diffusion would require us to track target searchers over several hundreds of nucleotides. Such long trajectories would imply very redundant scanning by Ago, and might therefore not be performed by the protein [**?**]. In an attempt to capture also short diffusive excursions [**? ? ? ? ?**], we utilized the high spatial resolution of smFRET [**?**].

The experimental procedure has been described in detail elsewhere [**? ?**], and we here restate only the core components. To trap any diffusive excursions for long enough to detect it (>100 ms), and have it complete before photobleaching (<700s), we design ss thymine (CbAgo [**?**]) and uracil (hAgo2, present study) repeats that contain two 3-nt targets and two 4-nt targets respectively (**Figures ??**A and **??**). In order to accurately determine whether the protein is binding to one target as opposed to the other, one of the traps is labeled with an acceptor fluorophore (Cy5), while the guide is labeled with the donor fluorophore

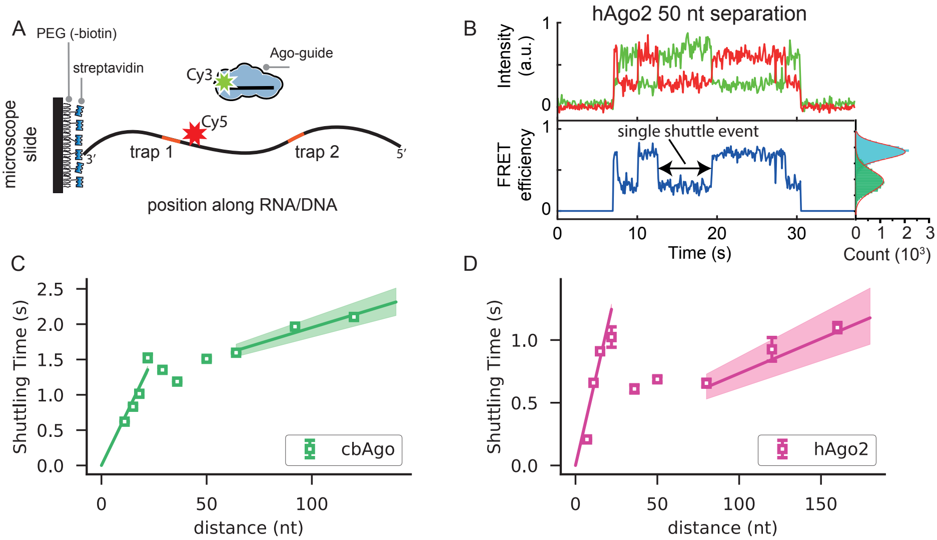


Figure 6.1: Single-molecule FRET experiment to probe lateral diffusion. (A) Schematic of assay. DNA/RNA constructs, containing the two trapping sequences (shown in red) are passivated to the microscope slide via a 3′ biotin-streptavadin linker and are labelled with the acceptor die. The Ago-guide complex is labelled with the donor die. **(B)** Representative trace for hAgo2 at a trap separation of 50nt. Top shows donor (green) and acceptor (red) signals. Bottom shows corresponding FRET efficiency and side panel shows histogram of all FRET efficiency values obtained for the population of molecules. **(C)** Shuttling time versus trapping distance (average ± sem) for CbAgo. Solid lines represent linear fits to data points at 11 nt ,15 nt ,18 nt ,22 nt (initial slope) and 64 nt ,92 nt ,120 nt (final slope). Shaded regions represent 95% confidence interval obtained using bootstrapping (see Methods). **(D)** Same as C for hAgo2. Data points at 7 nt ,11nt ,15 nt (initial slope) and 80 nt ,120 nt ,160 nt (final slope) are used for linear fits.

(Cy3) (**Figures ??**A and **??**). High FRET efficiency is observed when the protein binds to the site in close proximity of the acceptor dye, whereas lower FRET efficiency is obtained when Ago is trapped at the target far away from the dye (**Figure ??**B). To reduce the background fluorescence, traces were recorded using total internal reflection (TIRF) microscopy.

6.2.2. Ago slides over short distances

As shown in **Figure ??**B, the FRET efficiency shifts almost instantaneously between those corresponding to the two trap locations. Though smFRET solves the problem of spatial resolution, the total time spent diffusing now seems to have fallen below our time resolution (30-100ms). In a recent paper [**?**] we showed both experimentally and theoretically that for small trap separations, the average shuttling time is directly proportional to the trap separation

$$T_{\text{shuttle}}(d_{\text{trap}} < 25\text{nt}) \approx d_{\text{trap}}\tau_{\text{trap}} \tag{6.1}$$

with τ_{trap}^{-1} being the one-sided escape rate from the trapping sequence. The linear increase in shuttling time with trap separation is consistent with Ago performing rapid lateral diffusion (undetected), with numerous escape and re-trapping events before eventually making it across to the other trap (**Figure ??**A). In **Figure ??**C we show data for CbAgo [**?**], and in

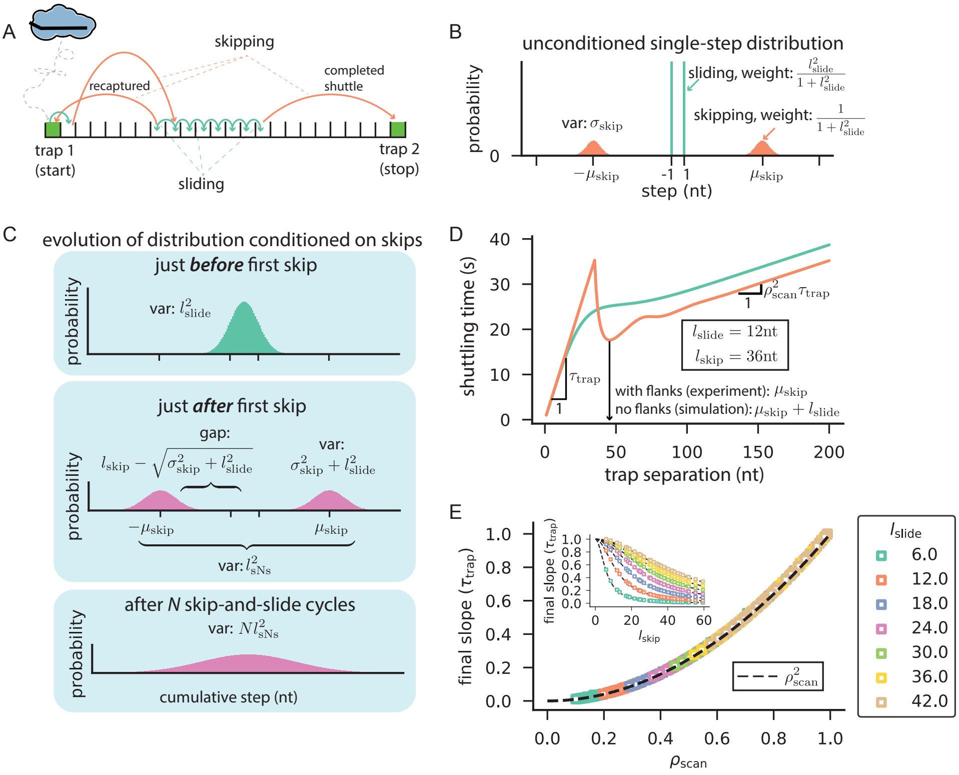


Figure 6.2: Modeling skip-and-slide search (shuttling events). (A) Schematic of shuttling event. Starting from the leftmost trap, the protein uses a combination of single-nucleotide steps (sliding) and larger steps (skipping) to reach the opposite trap after possibly getting recaptured at the initial traps several times. **(B)** single-step distribution of random walk defining our model. The protein either slides to a neighboring site or skips to sites located at $\pm(\mu_{\mathrm{skip}} \pm \sigma_{\mathrm{skip}})$. **(C)** Distribution of visited sites conditioned on skips. (top) The protein covers a rms distance l_{slide}^2 between consecutive skips. (middle) The first skip takes the protein μ_{skip} away (in either direction) with an uncertainty of σ_{skip} in the landing site. (bottom) Repeated skip-and-slide (sNs) cycles result in a distribution that resembles a simple random walk (top panel) with an adjusted effective step length of l_{sNs}. **(D)** Representative numerical solutions (**S.I.**) for shuttling time versus trapping distance. **(E)** Final slope versus scanning density. Inset shows equivalent versus skipping length (see **S.I.** for values in parameter sweep).

Figure ??D we confirm that the initial proportionality (**Equation ??**) reported for CbAgo also holds for hAgo2 (new data).

6.2.3. Ago uses a mixture of skipping and sliding over larger distances

As the distance between traps grows beyond the initial linear regime, the shuttling time drops, before it eventually settles into a gentler linear increase over large trap separations (**Figures ??**C and D). The drop in shuttling time suggests that a new avenue for traversing the gap between traps has opened up, while the shuttling time's eventual linearity with regard to trap separation suggests that also this avenue is governed by lateral diffusion and repeated re-trapping to the original trap, before reaching the second trap.To explain the linear long-range behavior, we consider the fact that CbAgo has previously been shown to bypass both protein roadblocks and secondary structures [**?**]. Exactly how such obstacles are traversed is not fully understood, but it is clear that bases would be skipped (i.e. not interrogated) in any process able to bypass roadblocks, and we will therefore simply refer to this process as skipping.

In **Figure ??**A we show a schematic of the skip-and-slide dynamics, and in **Figure ??**B we show the single step distribution such a random walker has within our model. In **Figure ??**C we show the cumulative step distribution conditioned on skipping. Measuring all lengths in nucleotides, Ago has diffused the average root-mean square (rms) distance l_{slide} after taking l_{slide}^2 sliding steps between consecutive skips (see **S.I.** for derivation). After having slid the l_{slide}^2 steps, Ago skips on average μ_{skip} nucleotides away in either direction, with a standard deviation of σ_{skip} nucleotides in the length of every skip (**Figures ??**B and C). In the **S.I.** we calculate the average shuttling time for such a system numerically using a master-equation formulation. In **Figure ??**D we show the resulting shuttling time for a fixed sliding length $l_{\text{slide}} = 12\text{nt}$, while the average skip distance and its standard deviation is either $\mu_{\text{skip}} = 36\text{nt}$ and $\sigma_{\text{skip}} = 0\text{nt}$ (green curve) or $\sigma_{\text{skip}} = 36\text{nt}$ and $\mu_{\text{skip}} = 0\text{nt}$ (orange curve). Both have the same rms skipping length, $l_{\text{skip}} = \sqrt{\mu_{\text{skip}}^2 + \sigma_{\text{skip}}^2} = 36\text{nt}$, with the $\sigma_{\text{skip}} = 0\text{nt}$ case representing skips of definite length that take the protein to a location not reachable in a single round of sliding ($l_{\text{skip}} \gg l_{\text{slide}}$). Contrarily, the protein may (likely) skip to a site already interrogated when $\mu_{\text{skip}} = 0\text{nt}$ – depleting the 'gap' shown in the middle panel of **Figure ??**C causes the distributions shown in the middle panel to overlap with that of the top panel. We note a clear resemblance of our numerical solutions to the empirical curves (**Figures ??**C and D), including the possibility of non-monotonic behavior when the skip length distribution is tight enough that there is a central gap in the cumulative step distribution just after the first skip (middle panel **Figure ??**C).
From the central-limit theorem it follows that the distribution of Ago positions after repeated skip-and-slide (sNs) cycles will approach that of simple diffusive motion with average mean squared step length $l_{\text{sNs}} = \sqrt{l_{\text{slide}}^2 + l_{\text{skip}}^2}$ between each unbinding cycle (bottom panel **Figure ??**C), where $l_{\text{skip}}^2 = \mu_{\text{skip}}^2 + \sigma_{\text{skip}}^2$ is the variance added to the cumulative translocation by one skip. In the **S.I.** we use a description conditioned on skips to construct scaling arguments showing that for large trap separations (**Figure ??**D)

$$T_{\text{shuttle}}(d_{\text{trap}} \gg l_{\text{sNs}}) \approx \text{const.} + \rho_{\text{scan}}^2 \tau_{\text{trap}} d_{\text{trap}} \text{ with } \rho_{\text{scan}} = {}^{l_{\text{slide}}}/_{l_{\text{sNs}}} \tag{6.2}$$

Here we have introduced the scanning density ρ_{scan} as the fraction of unique bases interrogated by Ago within a single skip-and-slide cycle. Having used our numerical approach to obtain $T_{\text{shuttle}}(d_{\text{trap}})$ curves for a wide range of l_{slide}, μ_{skip} and σ_{skip} (**S.I.**), the resulting final slopes from those curves indeed coincide with the derivative of **Equation ??**, thereby validating our scaling arguments (**Figure ??**E).

6.2.4. Ago skips straight into the second trap for intermediate trap separations

In between the two linear regimes, the shuttling time varies non-monotonically (**Figures ??**C,D and **??**D). At short distances, when only sliding, the protein's motion is well described by a simple random walk, with consecutive steps being uncorrelated (**Equation ??**). Using the scaling arguments leading up to **Equation ??**, a similar uncorrelated motion over segments of length l_{sNs} is expected at large trap separations. Although we expect said scaling arguments to fail (i.e. ignoring the constant in **Equation ??**) within the intermediate (non-monotonic) regime, preventing us from estimating the corresponding shuttling times, we can still estimate the trap separation at which we expect a local minimum shuttling time.
If the trap is not the outermost sequence on the construct, as is the case in our experiment (**Figure ??**), the initial sliding induces no average shift in position, and it stands to reason that the local minimum in shuttling times appears at a trap separation μ_{skip}, from where Ago typically slides straight into the second trap after the first skip. Below, we shall use this reasoning to estimate l_{slide} and l_{skip} from the data. Note that our numerical calculations have been performed for traps placed as the most outer sequence on the construct. For such a system Ago drifts an approximate distance l_{slide} towards the other trap before skipping, which is why **Figure ??**D shows a curve with its minimum around a trap separation of $\mu_{\mathrm{skip}} + l_{\mathrm{slide}} = 48\mathrm{nt}$ (orange curve).

6.2.5. Ago skips over two thirds of all bases

Applying the above arguments to our experimental data, we estimate the trapping time τ_{trap} by fitting **Equation ??** to the initial linear part of the shuttling time dependence on trap distance (left most line in **Figures ??**C and D, $\tau_{\mathrm{trap}} = 0.062 \pm 0.003\mathrm{s}$ for CbAgo and $\tau_{\mathrm{trap}} = 0.057 \pm 0.002\mathrm{s}$for hAgo2)(see **Methods**). Next, we can determine the scanning density ρ_{scan} by fitting **Equation ??** to the final linear part of the data (right most line in **Figures ??**C and D). The resulting scanning densities ($\rho_{\mathrm{scan}} = 0.38 \pm 0.03$ for CbAgo and $\rho_{\mathrm{scan}} = 0.31 \pm 0.04$ for hAgo2) indicate that only approximately one in three bases are checked by Ago while moving along its substrate.

We can further give rough estimates of the sliding distance and skip length as follows. As we see a dip in the shuttling time we know that skipping can only be a viable avenue of translocation above a certain trap separation, and thus there should be a gap in the position distribution of a skip-and-slide cycle just after the first skip (middle panel **Figure ??**C). For there to be a substantial gap in this distribution we need a clear separation between the distributions shown in the first two panels of **Figure ??**C. In mathematical terms, $\sigma^2_{\mathrm{skip}} + l^2_{\mathrm{slide}} \ll \mu^2_{\mathrm{skip}}$, implying that $l_{\mathrm{sNs}} \approx l_{\mathrm{skip}} \approx \mu_{\mathrm{skip}}$, and that the dip visible in the shuttling time (**Figures ??**C and D) essentially reports on this quantity. With a dip for both systems occurring around trap-separations of 30 nt, this implies a skipping distance of around $l_{\mathrm{skip}} \approx 30\mathrm{nt}$. With a scanning density of a third, this skip distance in turn suggests that both sliding distances are around $l_{\mathrm{slide}} \approx 10\mathrm{nt}$, or equivalently, Ago takes around 100 sliding steps between skips.

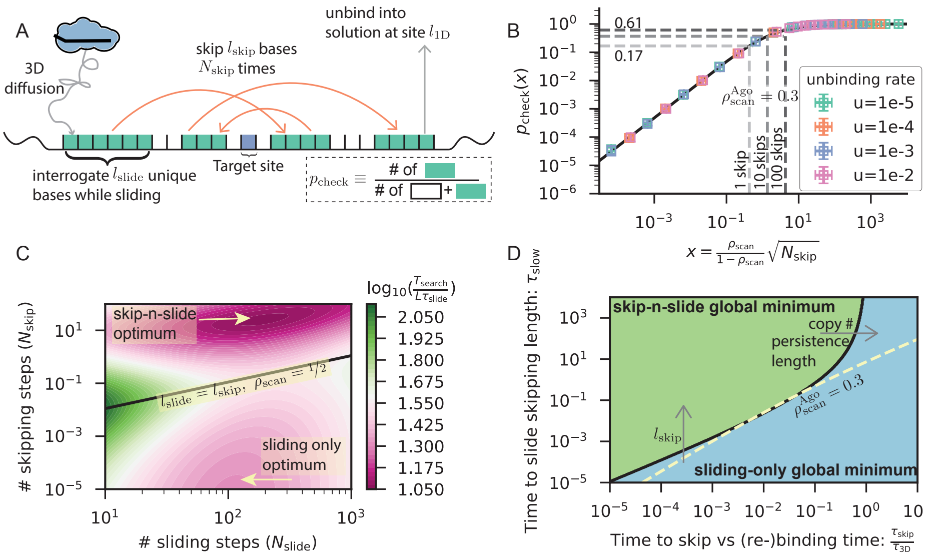


Figure 6.3: Optimal search times. (A) Schematic single search round. In search of the unique target, the protein uses a combination of skipping and sliding along the substrate before it unbinds into solution and must perform 3D diffusion before it can return. Only sites slid past (at least once) are interrogated (green), resulting in a probability p_{check} to interrogate a particular site. **(B)** Comparison of $p_{\mathrm{check}}(x)$ (solid line, Equation 3) to Monte Carlo simulations (symbols) (details given in Methods). Dashed lines indicate Argonauts ($\rho_{\mathrm{scan}} \approx 0.3$) that typically skip once (light grey), 10 (dark grey), and 100 times (black) before unbinding. **(C)** Search time versus N_{slide} and N_{skip}. Region above the solid line represents sparse scanning ($\rho_{\mathrm{scan}} < 0.5$), while the region below it represents dense scanning ($\rho_{\mathrm{scan}} > 0.5$). **(D)** Phase-diagram showing when $T^{\mathrm{sNs}}_{\mathrm{search}} < T^{\mathrm{sliding}}_{\mathrm{search}}$. Dashed line represents the constant scanning density of 0.3 (approx. the value estimated for both Ago). Arrows represent directions of increasing l_{skip}, protein copy number (concentration) and substrate persistence length.

6.2.6. The total search time

Having shown that both hAgo2 and CbAgo skip over a significant number of bases—about double the number of bases it actually scans in any skip-and-slide cycle—we now turn to the question why both Argonaute – from different kingdoms of life – behave so similarly. Under what conditions does skipping speed up a protein's search for a single target in the genome or mRNA pool? To answer this question, we now theoretically consider what combinations of the number and length of skipping and sliding steps – and thereby scanning density – lead to minimal overall search times.

We consider a target searcher that after diffusing through solution, binds its substrate randomly and non-specifically to perform a lateral excursion consisting of both skipping and sliding before unbinding (or finding the target). In a lateral excursion that ends with unbinding, we take the protein to undergo an average of N_{skip} skips, and N_{slide} slides. Note that N_{slide} does not equal the previously defined l^2_{slide}, as the latter is the number of sliding steps between consecutive skips, while the former equals l^2_{slide} multiplied by the number of skips prior to unbinding (see **S.I.**). To estimate the total time to find the target, we first determine the average number N_{rnd} of search rounds ('rnd') (binding-skip-and-slide-unbinding) needed before the target is found, and then the average time T_{rnd} of each search round [? ? ?]. In what follows, we express both N_{rnd} and T_{rnd} for target searchers using a mixture of skipping and sliding corresponding to a scanning density ρ_{scan}, after which we shall proceed to minimize the search time in terms of the frequency of skipping and sliding steps taken. To properly model the skip-and-slide process between unbinding events, we must cover the scenario presented in **Figure ??**A: even though the target sits in between the binding and unbinding locations, it might still be skipped over. In the **S.I.** we show that the average fraction of bases checked at least once over the rms lateral diffusion distance $l_{1\text{D}} = \sqrt{N_{\text{skip}}} l_{\text{sNs}}$ between binding and unbinding can be estimated using the scanning density and the typical number of skips prior to unbinding as (see **Figures ??**A and B)

$$p_{\text{check}}(x) = 1 - \frac{\log(1+2x)}{2x}, \text{ with } x = \frac{\rho_{\text{scan}}}{1-\rho_{\text{scan}}}\sqrt{N_{\text{skip}}} \tag{6.3}$$

The total number of checked sites at a fixed scanning density increases with increasing number of skips per binding event. The logic being that an increased number of skips allows for repeated rescanning of the same region of DNA sites, with the protein every time interrogating about ρ_{scan} of these sites. **Figure ??**B shows that if the Argonaute proteins ($\rho_{\text{scan}} \approx 0.3$) are to skip on average 100 times before unbinding, they still interrogate only about 60% of all sites spanned within its lateral excursion (dashed lines). Hence, after correcting for repeated scanning due to skipping, Ago likely still leaves a significant portion of the RNA/DNA unseen. We validated **Equation ??** (solid line in **Figure ??**B) using Monte Carlo simulations (colored data points, **Methods**).

Each lateral diffusion event checks on average $p_{\text{check}} l_{1\text{D}}$ distinct bases, and with a single target on a substrate of L nucleotides, it will take on average $N_{\text{rnd}} = {}^{L}/_{p_{\text{check}} l_{1\text{D}}}$ cycles before the target is found.

Each search round can be split between base interrogation through 1D lateral diffusion and 3D diffusion through solution. The 1D lateral diffusion time $\tau_{1\text{D}} = T_{\text{slide}} + T_{\text{skip}}$ can further be split into the total time spent interrogating off-targets after a sliding step $T_{\text{slide}} = N_{\text{slide}}\tau_{\text{slide}}$, and the total time spent completing skips and interrogating the landing site $T_{\text{skip}} = N_{\text{skip}}\tau_{\text{skip}}$. The timescales for interrogating off-targets after a sliding event τ_{slide}, executing skips τ_{skip} (including the time to interrogating the site of arrival), and executing excursions into solution $\tau_{3\text{D}}$ (including the time to interrogating the site of binding), together with the average number of rounds to find the target, leads us to the total search time

$$T_{\text{search}} = T_{\text{rnd}} N_{\text{rnd}} = (\overbrace{\underbrace{N_{\text{slide}}\tau_{\text{slide}}}_{T_{\text{slide}}} + \underbrace{N_{\text{skip}}\tau_{\text{skip}}}_{T_{\text{skip}}}}^{\tau_{1\text{D}}} + \tau_{3\text{D}}) \frac{L}{l_{1\text{D}} p_{\text{check}}} \tag{6.4}$$

We will seek the minima of the search time, but before proceeding we must consider what variables evolution could act upon to create a balance between skipping, sliding, and unbinding.

From the definition of the microscopic timescales we immediately have $\tau_{\text{skip}}, \tau_{3\text{D}} > \tau_{\text{slide}}$ as the sliding motion itself costs negligible time by assumption, and both skipping and excursions into solution are ended by interrogating the base at arrival (τ_{slide}). Further, we only ever expect to find an optimum with a balance between skipping and unbinding when the time to complete a skip is shorter than the time to return from solution. If returning from solution would be faster than completing a skip, skipping would always be eliminated and unbinding favored because it has both lower redundancy and is completed quicker. Further, decreasing any of the microscopic timescales associated with different search modes will clearly speed it up. Therefore, we assume these times to already be reduced as far as possible, and ordered as $\tau_{\text{3D}} > \tau_{\text{skip}} > \tau_{\text{slide}}$.

Apart from the three microscopic timescales, there are three more independent parameters evolution could act upon. These are the total number N_{skip} of skips N_{skip} in one search round, the number N_{slide} of off-targets checked after sliding in one search round, and the rms skip distance l_{skip} (or equivalently N_{skip}, N_{slide} and ρ_{scan}, see **S.I.**). Increasing only the rms skipping distance l_{skip} will always reduce the scanning redundancy, and so will always reduce the search time. Since we observe skips of finite length, we also assume these to be externally limited, and take also l_{skip} to be fixed. We are left with two independent parameters, and in **Figure ??**C we plot the search time as a function of N_{skip} and N_{slide} when $l_{\text{skip}} = 30\text{nt}$ and $\tau_{\text{3D}} = 10\tau_{\text{skip}} = 100\tau_{\text{slide}}$.

Minimization of the search time over our remaining two independent variables – the number of skips N_{skip} and the scanning density ρ_{scan} (defined in **Equation ??**) – results in two conditions that need to be satisfied at any optimum (see **S.I.**). We present the general conditions in the **S.I.**, and here present solutions valid in regimes of both high and low scanning densities to determine when skip-and-slide search, of the kind observed for Ago, is favored.

6.2.7. Sliding is optimal for scanning densities above 1/2

One local minimum exists in the densely scanned region ($1 - \rho_{\text{scan}} \ll 1/2$) and corresponds to the protein using sliding as its only lateral diffusion mode, eliminating skips entirely. The minimum is defined by, $\rho_{\text{scan}}^{\text{sliding}} = 1$, and (see **S.I.**)

$$N_{\text{skip}}^{\text{sliding}} = 0,\ N_{\text{slide}}^{\text{sliding}} = \frac{\tau_{\text{3D}}}{\tau_{\text{slide}}} \Rightarrow T_{\text{slide}} = \frac{1}{2}T_{\text{rnd}} \tag{6.5}$$

This minimum corresponds to the known minimum when a priori assuming that there are no skips [? ?]. Namely, the protein spends half its time diffusing through solution and the other half of the time sliding (the rightmost identity in **Equation ??** is equivalent to $\tau_{\text{1D}} = \tau_{\text{3D}}$). The search time at this minimum equals (see **S.I.**)

$$T_{\text{search}}^{\text{sliding}} = 2L\sqrt{\tau_{\text{slide}}\tau_{\text{3D}}} \tag{6.6}$$

The non-skip minimum is the only minimum in the densely scanned regime ($\rho_{\text{scan}} > 1/2$) (**Figure ??**C, minimum coinciding with horizontal axis), suggesting that it might be hard to evolve away from the it by incremental steps.

6.2.8. A mix of skipping and sliding is optimal for scanning densities below 1/2

For the skipping to be beneficial, skips must be large enough ($l_{\text{skip}} \gg l_{\text{slide}}$ or equivalently $\rho_{\text{scan}} \ll 1/2$) to get the system beyond the barrier visible in **Figure ??**C. In the **S.I.** we show that after recognizing

$$\tau_{\text{slow}} = \tau_{\text{slide}} l_{\text{skip}}^2 \tag{6.7}$$

as the time needed to traverse the length of a skip purely through sliding (diffusion with 1 nt steps) – a measure of the added benefit of using skipping – we obtain the location of the skip-n-slide optimum corresponding to a scanning density of $\rho_{\text{scan}}^{\text{sNs}} = l_{\text{skip}}^{-1}\sqrt{N_{\text{slide}}^{\text{sNs}}/N_{\text{skip}}^{\text{sNs}}} < 0.5$, with (see **S.I.**)

$$N_{\text{slide}}^{\text{sNs}} = \frac{\tau_{\text{3D}}}{\tau_{\text{slide}}}\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{\frac{1}{3}}, \; N_{\text{skip}}^{\text{sNs}} = \frac{\tau_{\text{3D}}}{\tau_{\text{skip}}}\left(1+\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{\frac{1}{3}}\right) \Rightarrow T_{\text{skip}} = \frac{1}{2}T_{\text{rnd}} \tag{6.8}$$

Note the final identity shown in **Equation ??** says that at the skip-and-slide optimum, the protein spends half of its time skipping, and the other half on a combination of sliding and diffusing through solution. In agreement with experimental studies [**? ?**], this indicates the protein spends more time diffusing along the DNA then it does through solution ($\tau_{\text{1D}} > \tau_{\text{3D}}$). The search time at this skip-and-slide optimum equals (see **S.I.**)

$$T_{\text{search}}^{\text{sNs}} = \frac{2L\sqrt{\tau_{\text{skip}}\tau_{\text{3D}}}}{l_{\text{skip}}} \frac{\sqrt{1+\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{\frac{1}{3}}}}{p_{\text{check}}\left(\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{\frac{2}{3}}\right)} \tag{6.9}$$

6.2.9. Global optimal search strategy

As there are local minima in both the sparsely and densely scanned regions (**Equations ??** and **??**), the global optimal search strategy is defined by which of these two minima have the smallest search time. The condition for the slip-and-slide minimum being the global minimum ($T_{\text{search}}^{\text{sNs}} < T_{\text{search}}^{\text{sliding}}$) can be written as (see **S.I.**)

$$\frac{\tau_{\text{skip}}}{\tau_{\text{slow}}} < \frac{p_{\text{check}}^{2}\left(\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{\frac{2}{3}}\right)}{1+\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{\frac{1}{3}}} < 1 \tag{6.10}$$

Figure ??D shows the corresponding phase diagram – in $\left\{\tau_{\text{slow}}, \frac{\tau_{\text{skip}}}{\tau_{\text{3D}}}\right\}$-space – showing when the skip-and-slide minimum is the global minimum. We previously argued that if $\tau_{\text{3D}} < \tau_{\text{skip}}$ there will be no skip-and-slide minimum. Now we see that for $\tau_{\text{3D}} > \tau_{\text{skip}}$ we can always find an l_{skip} long enough that the skip-and-slide optimum is also the global optimum (upward arrow in **Figure ??**D). Logically, the skip-and-slide optimum is only preferred over the sliding-only one for $\tau_{\text{slow}} > \tau_{\text{skip}}$, indicating the typical return time of a skip may not exceed the time needed to cover the same distance by just sliding, and **Equation ??** gives the more stringent condition that must be satisfied.

We conclude by noting both of the Argonaute proteins considered above have $\rho_{\text{scan}} \approx 0.3$ (yellow dashed line in **Figure ??**D), putting the system above the line separating the sparse and dense scanning regimes (**Figure ??**C). Certainly, hAgo2 and CbAgo operate far from the sliding-only optimum, and, as we shall discuss further below, are working in the regime where the skip-and-slide optimum is found (crossing point **Figure ??**D).

6.3. Discussion

Site-specific DNA or RNA binding proteins must find a single sequence amongst megabase (prokaryotes) to gigabase (eukaryotes) pools of off-targets. Here we have shown that facilitated diffusion with a mixture of sliding (single-nucleotide steps) with frequent and large skips (multi-nucleotide steps) is capable of reducing the overall search time beyond using sliding by itself. Interestingly, pure sliding is a possible optimal strategy, and the search time for skips shorter than the sliding length is minimal only after eliminating skips entirely as their temporal cost is no longer accompanied by the benefit of visiting off-targets not encountered before (**Figure ??**C). Contrarily, skips greater than the sliding length reduce the probability of redundantly sampling off-targets, and we find another optimum where the search time is minimal if skips are used so frequently that the system spends half of the time skipping. We further showed how single-molecule FRET experiments (**Figure ??**) can be used to extract what we termed the scanning density, a measure of the fraction of bases directly interrogated during a skip-and-slide cycle (**Figure ??**). Our experiments performed on a prokaryotic (CbAgo) and eukaryotic (hAgo2) Argonaute revealed both to have scanning densities around 0.3 (**Figures ??**C and **??**D)—well within the sparse scanning regime (**Figure ??**C).

As shown in **Figure ??**C, the scanning densities of the Argonaute proteins are consistent with having skip-n-slide search as an optimal strategy. However, according to **Figure ??**D it appears at this the system just touches the separating line determining the global optimum. One might speculate what other factor, not taken into account in our modeling, could have driven Ago away from the sliding only optimum. As shown in reference [**?**], skips are needed to surpass roadblocks present on any physiological substrate. Typical 3'-UTR substrates are 40-80% with proteins [**?**] and about one protein for every 30-100 nt is bound to cellular DNA [**? ?**]. We therefore hypothesize that if one limits the sliding length to be less than the typical separation between other (high affinity) binding proteins it to always be beneficial to include skips ($T_{\text{search}}^{\text{sNs}} < T_{\text{search}}^{\text{sliding}}$).

Based on our results, for a low scanning density to be preferred, the binding rate from solution should not exceed the return rate after skipping (**Figure ??**D). As binding rates scale linearly with concentrations (before reaching saturating levels), we thus expect binding proteins present at lower copy numbers to be prone to use more frequent skips (arrow in **Figure ??**D). For example, E.coli cells express about 1-10 copies of the lac repressor [**?**] and experiments have indeed seen signatures of a skipping-and-sliding mixture [**?**].
Instead of increasing (reducing copy number), a reduction in τ_{skip} is to be expected on more flexible substrates, such as single-stranded DNA or RNA. We therefore deem it likely that skip-n-slide search to also be used by sequence specific single-stranded binding proteins other than Argonaute, such as ribosomes searching for the transcription start site. We hope to motivate future experiments utilizing different DNA binding proteins to investigate whether they belong to the "sliding only" ($\rho_{\text{scan}} \gg 1/2$) or the "skipping-and-sliding" ($\rho_{\text{scan}} \ll \frac{1}{2}$) class (**Figure ??**C).

Within our analysis of the total search time we have decoupled the return time from a skip (τ_{skip}) from the average length thereof (l_{skip}). Hence, fixing the time, there is no penalty for ever increasing skipping distances. In fact, for large enough skipping distances we can always reach a situation where the skip-and-slide optimum is the global optimum (provided $\tau_{\text{skip}} < \tau_{3\text{D}}$)(**Figure ??**D). In our previous work [**?**] we demonstrated the duration of skips to be limited by the time needed to escape the bound site – rather than the time needed to find the distant location – justifying our assumption for Argonaute. However, skips limited by the rate of rebinding – for instance through diffusion – couple τ_{skip} to l_{skip} and we expect an optimal l_{skip} to exist. As we here focused on the coupling between search time and the experimentally measurable ρ_{scan}, we deem such an analysis beyond the scope of the presented research, but an interesting future direction.

A previous study [**?**] has pointed out that speeding up the lateral diffusion – by reducing the variation in binding strengths along the genome – comes at the cost of reducing the protein's specificity. The authors proposed that in order to overcome this apparent 'search-stability paradox' the protein must switch between two conformations – one with higher affinity (for specificity) and one with a lower one (for speed) – and detail the tight constrains on the binding energies for such a solution to exist [**? ?**](**Chapter ??**). Selected target searchers – including selected RNA guided nucleases [**? ? ? ? ?**] – indeed adopt multiple conformations during target interrogation [**? ? ?**]. The necessity for two protein conformations, however, arises from assuming the protein is only capable of sliding, thereby forcing the protein to sample every site along the genome. We hypothesize that using the different skip-and-slide scheme described here could provide a complementary/alternative route to being both fast and specific –allowing for wider spreads in binding energies – especially for proteins that are not known to exhibit multiple conformations.

The experiments performed here – together with our theoretical analysis – are in principle applicable to other DNA binding proteins. Proteins not guided by non-coding DNA/RNA should be labeled with the donor dye directly. Moreover, both Ago proteins examined here bind single-stranded nucleic acids, which have close to nucleotide persistent lengths [**?**] and thereby offer a clear possible mechanism of introducing frequent skips – Ago can skip to distant sequences as they can come close together in space. Yet, the presented analysis and experiment do not rely on such, and proteins binding double-stranded DNA – persistence lengths ~50 nt– can similarly be investigated for the presence of (presumably larger and less frequent) skips, without prior knowledge of a possible microscopic mechanism for skipping.

In conclusion, a search strategy combining skipping and sliding can significantly increase the rate of association to the cognate target – which is of critical importance for proper functioning of the cell – and Argonaute proteins adopt scanning densities consistent with their mixture being optimal.

6.4. Methods

6.4.1. Monte Carlo simulations for validating P_{check}

To test the validity of **Equation ??**, we set up Monte Carlo simulations (code written in Python). The proteins are assigned a unity step rate to either side, as well as an unbinding rate u. Hence in every move, the protein diffuses to one of its neighboring site with a probability $^{1}/_{2+u}$ and unbinds with a probability $^{u}/_{2+u}$. Before every move, the protein interrogates the site currently located at with a fixed probability of ρ_{scan}. Each of the 1000 runs ends when the protein unbinds. The corresponding value of x is evaluated using the distance between binding and unbinding sites (see definition of x above **Equation ??**). We estimate the value of p_{check} as the fraction of sites visited that are interrogated. Error bars in **Figure ??**B show 95% confidence intervals for both x and p_{check}. Simulations we repeated for in [10^{-10}, 10^{-9},10^{-8},...,10^{-2}, 0.9,0.8,...0.1], and u in [10^{-5},...,10^{-2}] as indicated in **Figure ??**B.

6.4.2. Bootstrapping for error estimation and based on smFRET data

Fitting the data from the tandem target assay to **Equation ??** provides the estimate of τ_{trap}. We bootstrapped the dwell time distributions acquired using the original tandem target assay (distances of 11 nt, 15 nt, 18 nt and 22 nt (CbAgo) and 7 nt, 11 nt, and 15 nt (hAgo2)). For each of the 10^5 bootstrap samples we calculated new values for the associated T_{shuttle}'s and repeated the fit to **Equation ??** to obtain an error estimate in the fitted value of the escape rate. In similar fashion, we used **Equation ??**, together with the estimate of τ_{trap} from the original dataset, to determine ρ_{scan} (distances of 64 nt, 92 nt and 120 nt (CbAgo) and 80 nt, 120 nt, 160 nt (hAgo2)). All analysis was performed with a custom code written in Python. Shaded areas in **Figures ??**C and D represent 95% confidence intervals.

6.4.3. protein purification

CbAgo was purified according to Hegge et al, 2019 [**?**]. hAgo2 was purified according to Chandradoss et al, 2015 [**?**].

6.4.4. Nucleic acid preparation

RNA constructs with a single amine-C6-uridine modification were ordered from STPharm. After labelling with Cy5 according to [**?**], the constructs were precipitated. The RNA constructs were subsequently annealed to a DNA splint (specific for RNA and U40 mer), a second DNA splint (for ligating U40 mers) and a U40 mer (in the ratio 1:2:3:3). After ligation with T4 RNA ligase II (NEB), the ligated constructs were run on a 10% PAGE. Different ligated populations are created through this process (for example, TGT- U40 or TGT-U40-U40 etc) and these are then excised from the gel and concentrated through ethanol precipitation. The concentrated and ligated RNA constructs were again annealed to a DNA construct and an RNA target with biotin on the 3' end. Ligation was again performed with T4 RNA ligase II. DNA oligos with a single amine-C6-thymine modification were ordered from ELLA Biotech GmbH and labeled in the same way as the RNA.

6.4.5. Sample preparation

Quartz slides were prepared according to [**?**]. Briefly, quartz slides were cleaned with detergent, sonicated and treated with acetone and subsequently KOH. Coverslips were directly sonicated with KOH. Piranha cleaning was done followed by treatment with methanol and incubation of (3-Aminopropyl)triethoxysilane (APTES) for both coverslips and quartz slides. PEGylation took place overnight and slides and coverslips were stored at -20 °C. Before single-molecule experiments, an extra round of PEGylation took place with MSPEG-4. The quartz slide was then assembled with scotch tape and epoxy glue and the chamber is flushed in T50 and 1% Tween-20 for >10min to further improve the surface quality of the single-molecule chambers [**?**]. Channels were thoroughly washed with T50 before adding in streptavidin (0.1 mg/mL) for 1 min. Subsequently, DNA or RNA was immobilized on the surface through biotin-streptavidin conjugation. 10 nM CbAgo or hAgo2 was incubated with 1 nM guide in (100 mM NaCl for CbAgo, 50 mM NaCl for hAgo2), 50 mM Tris, 1 mM Trolox, 0.8% glucose for 30 min. Lastly, glucose oxidase (0.1 mg/mL final conc.) and catalase (17 µg/mL final conc.) were added and introduced in the chamber.

6.4.6. Experimental setup

Single-molecule experiments were performed on a custom built inverted microscope (IX73, Olympus) using prism-TIRF and a 60X water immersion objective (UPLSAPO60XW, Olympus). The Cy3 dye was excited using a 532 nm diode laser Compass 215M/50mW, Coherent) and the Cy5 dye was excited using a 637 nm diode laser (OBIS 637 nm LX 140 mW). The scattered light was blocked by a 532 nm notch filter (NF03-532E-25, Semrock) and a 633 nm notch filter (NF03-633E-25, Semrock) after which the remaining signal from the fluophores was separated into two separate channels. Lastly, the light is projected on a EM-CCD camera (iXon Ultra, DU-897U-CS0-# BV, Andor Technology). Before each experiment, a reference movie was taken with the red laser to excite the Cy5 dyes on the nucleic acid molecules of interest. After that, a movie is taken with the green laser. The single-molecule experiments were taken at room temperature (20 $\pm$ 0.1 °C).

6.4.7. Analysis of raw data

The raw data was analysed using custom written code in IDL, where the reference movie is used to take into account only the regions of interest (i.e. the regions that contain a Cy5). The resulting time traces where further analysed in MATLAB (Mathworks) where the shuttling rates were extracted through the use of Hidden Markov software called ebFRET (http://ebfret.github.io/) and custom written code in Matlab.

6.5. Author contributions

M.K. and M.D. performed all theoretical analysis, T.J.C. and C.J. designed the experiments, and T.J.C. performed the experiments. I.M. provided the hAgo2 protein. M.K., T.J.C., M.D. and C.J. wrote the manuscript.

6.6. Acknowledgments

This work (M.K.) was supported by the Netherlands Organization for Scientific Research, as part of the Frontiers in Nanoscience program.

6.7. Supplemental Information

6.7.1. Determining shuttling times using a mixture of skipping and sliding

We here build a kinetic model for the lateral diffusion by target searching proteins capable of explaining the experimental data shown in **Figure ??**.

modeling skipping-and-sliding lateral diffusion

Given the protein can in principle (attempt to) bind any sequence along the DNA or RNA, we imagine binding sites to be a nucleotide apart. When bound to site i, the protein diffuses away (in either direction) at a rate

$$k_{\text{move}}(i) = \begin{cases} k_{\text{trap}} & \text{at trap} \\ k_{\text{ns}} & \text{at non-specific site} \end{cases} \tag{S6.1}$$

We assume the binding energy at the trap is significantly greater than at any non-specific site, with both still being significantly more stable than the unbound state. As a result, the (average) shuttling time measured in our *in vitro* experiments - the system contains two stronger binding traps and a limited amount of remaining off-targets - is governed by movements from the trap.

$$k_{\text{ns}} \gg k_{\text{trap}} \tag{S6.2}$$

Ignoring any temporal contribution from the non-specific sites reflects the lack of any directly observable FRET signal corresponding to the protein being at these locations (**Figure ??**). Furthermore, given the TIRF microscopy assay ensures we are tracking laterally diffusing proteins that did not unbind - proteins diffusing through solution move in and out of the evanescent field too fast to be detected - we shall ignore the protein's intrinsic unbinding rate at all sites for now - an assumption that is further justified by noting that typically more than 10 shuttle events occur prior to unbinding.

In every move, taking an average time of k_{move}^{-1}, the protein can either slide - step to its neighbors - or skip - step further.We let the rate to step away from site i still be set by **Equation ??** and assign a probability that such a step is of definite length $|l|$ (in nucleotides). Letting $\delta_{x,y}$ denote the Kronecker delta,

$$P(l, l_{\text{slide}}, l_{\text{skip}}) = \frac{n_{\text{slide}}(l_{\text{slide}})}{1 + n_{\text{slide}}(l_{\text{slide}})}\delta_{|l|,1} + \frac{1}{1 + n_{\text{slide}}(l_{\text{slide}})} s(|l|, l_{\text{skip}}) \tag{S6.3}$$

, with $\sum_{n>0} P(n) = 1$. The weight of a skip of length $|l|$ as a function of the typical skipping length l_{skip}, is denoted by $s(|l|, l_{\text{skip}})$. Further, n_{slide} is the typical number of sliding steps taken between two consecutive skips. Given a sliding step displaces the protein by a single nucleotide, the stochastic variable Δn_i representing the number of nucleotides moved during one such step follows

$$\Delta n_i = \begin{cases} +1 \text{ nt} & p = 1/2 \\ -1 \text{ nt} & 1 - p = 1/2 \end{cases} \tag{S6.4}$$

Hence, the mean squared displacement after n_{slide} of such steps equals

$$\begin{aligned}(1\,\text{nt})^2\, l_{\text{slide}}^2 &= \langle \left(\sum_{i=1}^{n_{\text{slide}}} \Delta n_i \right)^2 \rangle \\ &= \sum_{i=1}^{n_{\text{slide}}} \sum_{j=1}^{n_{\text{slide}}} \langle \Delta n_i \Delta n_j \rangle \\ &= \sum_{i=1}^{n_{\text{slide}}} \langle (\Delta n_i)^2 \rangle + \sum_{i \neq j} \langle \Delta n_i \rangle \langle \Delta n_j \rangle \\ &= n_{\text{slide}} \times \langle (\Delta n_1)^2 \rangle \\ &= n_{\text{slide}}\, (1\,\text{nt})^2 \end{aligned} \tag{S6.5}$$

, where in the third line we have used the independence of individual steps. We define the 'sliding length', $l_{\text{slide}} = \sqrt{n_{\text{slide}}}$, as the typical number of nucleotides covered sliding between two consecutive skips - the rms displacement of a simple random walk with n_{slide} steps. Rewritten in terms of the now defined sliding length l_{slide}, the probability of taking a step of length $|n|$ reads

$$P(n, l_{\text{slide}}, l_{\text{skip}}) = \frac{l_{\text{slide}}^2}{1 + l_{\text{slide}}^2} \delta_{|n|,1} + \frac{1}{1 + l_{\text{slide}}^2} s(|n|, l_{\text{skip}}) \tag{S6.6}$$

The (effective) rate from i to j then equals

$$\kappa(i, j | l_{\text{slide}}, l_{\text{skip}}) = k_{\text{move}}(i) P(|i - j|, l_{\text{slide}}, l_{\text{skip}}) \tag{S6.7}$$

As we will show below, the behavior of the resulting shuttling times both at short and long distances is independent of the choice of the distribution s. Yet, all numerical results are obtained using

$$s(n, \mu_{\text{skip}}, \sigma_{\text{skip}}) = \int_{n-1/2}^{n+1/2} \left[G(n|\mu_{\text{skip}}, \sigma_{\text{skip}}) + G(n| - \mu_{\text{skip}}, \sigma_{\text{skip}}) \right] \mathrm{d}n \tag{S6.8}$$

with

$$G(x, \mu_{\text{skip}}, \sigma_{\text{skip}}) = \frac{1}{\sqrt{2\pi\sigma_{\text{skip}}^2}} e^{\frac{-(x-\mu_{\text{skip}})^2}{2\sigma_{\text{skip}}^2}} \tag{S6.9}$$

denoting the Gaussian distribution with average μ_{skip} and standard deviation σ_{skip}. Hence, the length of each skip is normally distributed, with a typical (rms) skipping length of

$$l_{\text{skip}} = \sqrt{\mu_{\text{skip}}^2 + \sigma_{\text{skip}}^2} \tag{S6.10}$$

numerical method to solve for shuttling time

Every shuttling event starts with the protein bound at one of the two trapping sites ($t = 0$) and ends the first time it reaches the other ($t = T_{\text{shuttle}}$), located d_{trap} sites away. Using the transition rates of **Equation ??**, letting $P_{\text{i}}(t)$ denote the probability for the protein to reside at site i at time t, and defining the vector

$$\vec{P}(t) = \left[P_1(t), P_2(t), ..., P_{d_{\text{trap}}-1}(t)\right]^T \tag{S6.11}$$

(for ease of notation we omit the sites flanking either trap $i < 1$ and $i > d_{\text{trap}}$, but note the approach mentioned here is applicable also if the traps are not the outermost sites on the construct)

the following set of Master Equations determine the evolution of the occupancies at every site during a shuttling event with the first trap at site 1 and the second at d_{trap}.

$$\frac{\partial \vec{P}}{\partial t} = -K\vec{P}(t) \tag{S6.12}$$

with the elements in rate matrix K given by

$$K_{ij} = \begin{cases} -\kappa(|j-i|, l_{\text{slide}}, l_{\text{skip}}) & \forall i \neq j \\ \sum_{i\neq j} \kappa(|i-j|, l_{\text{slide}}, l_{\text{skip}}) & \forall i = j \end{cases} \tag{S6.13}$$

The shuttle event starts with the protein located at the first trap,

$$P_1(0) = 1, \quad P_{\text{i}}(0) = 0 \; \forall i \neq 1 \tag{S6.14}$$

, and ends when the second trap is reached, whose corresponding outgoing rates are set to zero ($j = d_{\text{trap}}$ in **Equation ??**). The probability of completing a shuttle within the time interval $[\tau, \tau{+}\Delta t]$ should be proportional to the change in occupancy at the destination trap ($P_{d_{\text{trap}}}(\tau + \Delta t) - P_{d_{\text{trap}}}(\tau)$). Letting $p_{\text{shuttle}}(\tau)$ denote the probability density of completing the shuttle at time τ, ($p_{\text{shuttle}}(\tau)\Delta t = P_{d_{\text{trap}}}(\tau + \Delta t) - P_{d_{\text{trap}}}(\tau)$, for small enough Δt. Taking $\Delta t \to 0$, we recognize the rate of change of the second trap's occupancy ($\frac{\partial P_{d_{\text{trap}}}(t)}{\partial t}|_{t=\tau}$) as the instantaneous probability that the shuttling time equals τ ($p_{\text{shuttle}}(\tau)$). Denoting the basis vectors $\vec{p_{\text{j}}}$ as $\vec{p_0} = [1, 0, 0, ...0]^T$, $\vec{p_1} = [0, 1, 0, ...0]^T$, $\vec{p_2} = [0, 0, 1, ...0]^T$ and so on, the shuttle times are distributed as

$$\begin{aligned} p_{\text{shuttle}}(\tau) &= \frac{\partial P_{d_{\text{trap}}}(t)}{\partial t}\Big|_{t=\tau} \\ &= -\sum_{j\neq d_{\text{trap}}} \frac{\partial P_j(t)}{\partial t}\Big|_{t=\tau} \\ &\equiv -\sum_{j\neq d_{\text{trap}}} \vec{p}_{\text{j}}^{\,T} \frac{\partial \vec{P}(t)}{\partial t}\Big|_{t=\tau} \\ &= +\sum_{j\neq d_{\text{trap}}} \vec{p}_{\text{j}}^{\,T} K\vec{P}(\tau) \\ &= +\sum_{j\neq d_{\text{trap}}} \vec{p}_{\text{j}}^{\,T} Ke^{-K\tau}\vec{P}(0) \end{aligned} \tag{S6.15}$$

In the second line we have used that any additional occupancy at the trap must come from somewhere else on the RNA/DNA ($P_{d_{\text{trap}}}(t) = 1 - \sum_{j\neq d_{\text{trap}}} P_{\text{j}}$). The next lines makes use of **Equation ??** together with the basis vectors to write the elements of $\vec{P}$ as its projections, and the Master Equation, **Equation ??**, to work in the rate matrix K and its matrix exponential.The desired average shuttling time (T_{shuttle}) is the first moment of the distribution $p_{\text{shuttle}}(\tau)$,

$$\begin{aligned} T_{\text{shuttle}}(d_{\text{trap}}) &= \int_0^\infty \tau p_{\text{shuttle}}(\tau)\mathrm{d}\tau \\ &= \int_0^\infty \tau \sum_{j\neq d_{\text{trap}}} \vec{p}_{\text{j}}^{\,T} Ke^{-K\tau}\vec{P}(0)\mathrm{d}\tau \\ &= \sum_{j\neq d_{\text{trap}}} \vec{p}_{\text{j}}^{\,T} \left(\int_0^\infty \tau Ke^{-K\tau}\mathrm{d}\tau\right)\vec{P}(0) \\ &= \sum_{j\neq d_{\text{trap}}} \vec{p}_{\text{j}}^{\,T} K^{-1}\vec{P}(0) \end{aligned} \tag{S6.16}$$

Using the values of l_{slide}, μ_{skip} and σ_{skip} (thereby knowing l_{skip} via **Equation ??**) and the distance between traps d_{trap}, we construct the rates in **Equation ??**, build the matrix K, invert it and compute $T_{\text{shuttle}}(d_{\text{trap}})$ as the inner product shown in **Equation ??**. Note that if the trap located at d_{trap} is not the outermost binding site on the construct, **Equation ??** is still valid after substituting matrix K for the sub-matrix with its d_{trap}-th row and column removed.

6.7.2. Shuttling times scales with square of scanning density at large trap separations

Given movements along the non-specific parts of the substrate occurred too fast to be observed, T_{shuttle} should be proportional to the time needed to escape the initial trap towards the region in between traps ($\tau_{\text{trap}} = k_{\text{trap}}^{-1}$) multiplied by the number of re-trapping events.

$$T_{\text{shuttle}}(d_{\text{trap}}) = n_{\text{return}}\tau_{\text{trap}} \tag{S6.17}$$

After sufficient rounds of skipping and sliding, the protein's excursion is well described by a random walk with basic step length (**Figure ??**, 'sNs':'skip-N-slide'):

$$l_{\text{sNs}} = \sqrt{l_{\text{slide}}^2 + l_{\text{skip}}^2} = \sqrt{l_{\text{slide}}^2 + \mu_{\text{skip}}^2 + \sigma_{\text{skip}}^2} \tag{S6.18}$$

The protein slides - covering l_{slide} nucleotides - before skipping to the next segment of length l_{sNs}. For this coarse-grained system, we once again expect the escaping of the trap to be rate limiting, resulting again in a linear increase of the shuttling time with inter-trap distance, similar to the case of diffusion purely by sliding (**Equation ??**),

$$T_{\text{shuttle}}(d_{\text{trap}}) = \text{const.} + \hat{n}_{\text{return}}\tau_{\text{trap}} \tag{S6.19}$$

Here we are concerned only with $T_{\text{shuttle}}(d_{\text{trap}})$'s scaling with d_{trap}, for which it is only the term proportional to τ_{trap} that has to be taken into account. In the coarse-grained system

$$\begin{aligned}\hat{n}_{\text{return}} &= (\text{\# returns to segment that contains the first trap})\\ &\times (\text{\# returns to trap when in first segment})\\ &\equiv \hat{n}_{\text{segment}} \times \hat{n}_{\text{retrap}}\end{aligned} \tag{S6.20}$$

To get the average number of re-entries to the first segment we must derive its corresponding probability. First, given a skip translocates the protein to an adjacent segment of l_{sNs} nucleotides, and l_{slide}^2 steps are taken within each segment

$$\rho_{\text{scan}} = \frac{l_{\text{slide}}}{l_{\text{sNs}}} = \frac{l_{\text{slide}}}{\sqrt{l_{\text{slide}}^2 + l_{\text{skip}}^2}} \tag{S6.21}$$

denotes the typical fraction of interrogated sites along the substrate, or 'scanning density'. In other words, any particular site within a l_{sNs}-long region of DNA/RNA has a probability of ρ_{scan} to be interrogated prior to the protein moving beyond this segment. Equivalently, the protein visits a segment without checking (all) the sites within it with a probability of $1 - \rho_{\text{scan}}$. Next, let $P_{\text{shuttle}}(\hat{d})$ denote the probability of traversing/shuttling across $\hat{d}$ segments without entering the previous segment. We shall derive $P_{\text{shuttle}}(d)$ below. Having entered the first of the $\hat{d}_{\text{trap}} = {}^{d_{\text{trap}}}/_{l_{\text{sNs}}}$ segments that lie between the traps, the probability of returning to the segment that contains the initially bound trap equals (**Figure ??**).

$$\begin{aligned}P_{\text{segment}} &= \left(1 - P_{\text{shuttle}}(\hat{d}_{\text{trap}})\right)\\ &\quad + P_{\text{shuttle}}(\hat{d}_{\text{trap}}) \sum_{m=0}^{\infty} \left((1-\rho_{\text{scan}})\left(1 - P_{\text{shuttle}}\right)\right)^m (1-\rho_{\text{scan}})\, P_{\text{shuttle}}(\hat{d}_{\text{trap}})\end{aligned} \tag{S6.22}$$

The first term is the probability of immediately going back to the segment the protein started from, while the sum accounts for the probability of all paths that reach the segment that contains the second trap, do not get captured by it, and eventually return back to the first trap (**Figure S2**). For instance, the $m = 0$ term ($P_{\text{shuttle}}(1 - \rho_{\text{scan}})P_{\text{shuttle}}$) represents the path that walks to the opposite side of the construct, does not interrogate the final trap and walks back across the construct to arrive back at the segment with the initially bound trap.

Using a similar type of 'path counting', we find the probabilities P_{shuttle} and $P_{\text{no shuttle}} = 1 - P_{\text{shuttle}}$, for a given inter-trap distance $\hat{d}_{\text{trap}}$ to equal (**Figure ??**)

$$P_{\text{no shuttle}}(\hat{d}_{\text{trap}}) = \sum_{m=0}^{\infty} \left(\frac{1}{2} \left(1 - P_{\text{shuttle}}(\hat{d}_{\text{trap}} - 1)\right) \right)^m \frac{1}{2} \tag{S6.23}$$

$$P_{\text{shuttle}}(\hat{d}_{\text{trap}}) = \sum_{m=0}^{\infty} \left(\frac{1}{2} \left(1 - P_{\text{shuttle}}(\hat{d}_{\text{trap}} - 1)\right) \right)^m \frac{1}{2} P_{\text{shuttle}}(\hat{d}_{\text{trap}} - 1) \tag{S6.24}$$

- from which we can write the recurrence relation

$$P_{\text{shuttle}}(\hat{d}_{\text{trap}}) = P_{\text{no shuttle}}(\hat{d}_{\text{trap}}) P_{\text{shuttle}}(\hat{d}_{\text{trap}} - 1) \tag{S6.25}$$

The above can be re-written as

$$P_{\text{shuttle}}(\hat{d}_{\text{trap}}) = \frac{P_{\text{shuttle}}(\hat{d}_{\text{trap}} - 1)}{P_{\text{shuttle}}(\hat{d}_{\text{trap}} - 1) + 1} \tag{S6.26}$$

, which subjected to the boundary condition $P_{\text{shuttle}}(1) = 1$ - signifying that if the traps are placed in adjacent segments, the shuttle is complete once the protein escaped the trap for the first time - has the simple solution

$$P_{\text{shuttle}}(\hat{d}_{\text{trap}}) = \frac{1}{\hat{d}_{\text{trap}}} \tag{S6.27}$$

Given the probability of re-entering the first segment, the average number of times this occurs prior to eventually shuttling across equals

$$\hat{n}_{\text{segment}} = \sum_{n=0}^{\infty} n P_{\text{segment}}^{n} (1 - P_{\text{segment}}) = \frac{P_{\text{segment}}}{1 - P_{\text{segment}}} \tag{S6.28}$$

Using **Equation ??** we find that the protein on average re-enters the segment with the initial trap

$$\hat{n}_{\text{segment}} = \frac{d_{\text{trap}}}{l_{\text{sNs}}} + \frac{l_{\text{sNs}}}{l_{\text{slide}}} - 2 \tag{S6.29}$$

times prior to completing the shuttling event. Once arrived back within the first segment, we must count the (average) number of times the protein gets recaptured by the actual trap

($\hat{n}_{\text{retrap}}$). Assuming sufficient ’skip-and-slide cycles’ have taken place, the protein’s position is uniformly spread throughout the l_{sNs}-long segment (**Figure ??**C). Hence, every step taken within the segment has a probability of $1/l_{\text{sNs}}$ to lead to the trap. Given there are typically $n_{\text{slide}} = l^2_{\text{slide}}$ steps taken prior to a skip (that moves the protein outside of the l_{sNs}-long region),

$$\hat{n}_{\text{retrap}} = \frac{n_{\text{slide}}}{l_{\text{sNs}}} = \frac{l^2_{\text{slide}}}{l_{\text{sNs}}} \tag{S6.30}$$

Taken together, **Equations ??** and **??** - by virtue of **Equation ??**:

$$\hat{n}_{\text{return}} = \frac{l^2_{\text{slide}}}{l_{\text{sNs}}} \times \left[\frac{d_{\text{trap}}}{l_{\text{sNs}}} + \frac{l_{\text{sNs}}}{l_{\text{slide}}} - 2\right] \equiv \text{const.} + \frac{l^2_{\text{slide}}}{l^2_{\text{sNs}}} d_{\text{trap}} \tag{S6.31}$$

Hence, when placed sufficiently far apart, the shuttling time (**Equation ??**),

$$T_{\text{shuttle}}(d_{\text{trap}}) = \text{const.} + \rho^2_{\text{scan}}\tau_{\text{trap}} d_{\text{trap}} = \left(\frac{1}{1 + \left(\frac{l_{\text{skip}}}{l_{\text{slide}}}\right)^2}\right) d_{\text{trap}} \tag{S6.32}$$

grows linearly with a slope that scales quadratically with the scanning density (**Equation ??**) from which we obtain the ratio between sliding and skipping lengths.

6.7.3. parameter sweep and estimation of slopes

To construct **Figure** 2E, we evaluate **Equation ??** for $l_{\text{slide}} \in$ [1 nt, 6 nt, 12 nt, 18 nt, 24 nt, 30 nt, 36 nt, 42 nt], $\mu_{\text{skip}} \in$ [0 nt,6 nt, 12 nt, 18 nt, 24 nt, 30 nt, 36 nt, 42 nt] and $\sigma_{\text{skip}} \in$ [0.01 nt, 6 nt, 12 nt, 18 nt, 24 nt, 30 nt, 36 nt, 42 nt]. The distance between traps varied from 1-250 nt. The values of l_{slide}, μ_{skip} and σ_{skip} where chosen such that at the largest trap separation of 250 nt the system is always in the regime for which we expect **Equation ??** to hold.
For every T_{shuttle} vs d_{trap} curve, we use the first two points (1 nt , 2 nt) to estimate τ_{trap} (**Equation ??**) and the final two points (249 nt, 250 nt), together with the estimate of τ_{trap}, to estimate ρ_{scan} (**Equation ??**).

6.7.4. Search time using skipping and sliding shows two optima

Here we connect the scanning density (ρ_{scan}) that we can extract from experiments to the time needed for a protein to locate a single target embedded within a larger pool of L binding sites. Following [**?**],

$$T_{\text{search}} = N_{\text{rnd}} T_{\text{rnd}} \tag{S6.33}$$

with T_{rnd} the (average) time each round of facilitated diffusion takes and N_{rnd} the number of such rounds (’rnd’) needed to find the target. As mentioned in the main text, we seek to find the minimum search time with respect to the number skips (N_{skip}) and slides (N_{slide}) within every round (binding - lateral diffusion - unbinding).

The length of a skip (l_{skip}), as well as the times to interrogate (slide past) a binding site (τ_{slide}), execute a skip and interrogate the landing site (τ_{skip}), and the time spent on 3D diffusion and interrogating the landing site (τ_{3D}) are all kept constant. The time per round consists of the time spent on the DNA performing lateral diffusion and the time spent in solution performing 3D diffusion.

$$T_{\text{rnd}} = \tau_{\text{1D}} + \tau_{\text{3D}} \tag{S6.34}$$

We further write the time spent on lateral diffusion as the time spent interrogating off-targets either by sliding or skipping,

$$\tau_{\text{1D}} = T_{\text{slide}} + T_{\text{skip}} \tag{S6.35}$$

For ease of calculation, we define the following variables with respect to which we have minimized the search time

$$x = \frac{\rho_{\text{scan}}}{1-\rho_{\text{scan}}}\sqrt{N_{\text{skip}}} \tag{S6.36}$$

$$y = \frac{\rho_{\text{scan}}}{1-\rho_{\text{scan}}} \tag{S6.37}$$

Written in terms of x and y (**Equations ??** and **??**), the total times spent either on sliding or skipping become

$$T_{\text{slide}} = N_{\text{slide}}\tau_{\text{slide}} = (x\delta l)^2\tau_{\text{slide}} \tag{S6.38}$$

$$T_{\text{skip}} = N_{\text{skip}}\tau_{\text{skip}} = \left({}^{x}\!/_{y}\right)^2\tau_{\text{skip}} \tag{S6.39}$$

Here we have introduced the variable $\delta l = l_{\text{sNs}} - l_{\text{slide}} = \frac{l_{\text{skip}}}{\sqrt{1+2y}}$ for ease of notation. To complete **Equation ??** we need the average number of search rounds (binding-lateral diffusion-unbinding) needed to locate a single target amongst L potential binding/target sites,

$$N_{\text{rnd}} = \frac{L}{l_{\text{1D}}p_{\text{check}}(x)} \tag{S6.40}$$

In here, we set the typical length of a lateral excursion to span l_{1D} sites, out of which a fraction $p_{\text{check}}(x)$ have been interrogated (slid past) at least once prior to unbinding (derivation shown below) (see **Figure 3A**). Further, l_{1D} represents the (rms) distance between binding and unbinding sites

$$\begin{aligned} l_{\text{1D}} &= \sqrt{N_{\text{slide}} + N_{\text{skip}}l_{\text{skip}}^2} \\ &= \sqrt{N_{\text{skip}}n_{\text{slide}} + N_{\text{skip}}l_{\text{skip}}^2} \\ &= \sqrt{N_{\text{skip}}l_{\text{slide}}^2 + N_{\text{skip}}l_{\text{skip}}^2} \\ &= \sqrt{N_{\text{skip}}}l_{\text{sNs}} = \left(\frac{y+1}{y}\right)x\delta l \end{aligned} \tag{S6.41}$$

In the second line of **Equation ??** we have rewritten the total number of sites visited through sliding as the product of the number of skip-n-slide cycles (N_{skip}) and the number of sliding steps between two skips (n_{slide}). The latter is related to the sliding length as we have defined it above ($l_{\text{slide}}^2 = n_{\text{slide}}$, **Equation ??**). In the last line, we recognize the rms length covered in a skip-n-slide cycle ($l_{\text{sNs}} = \sqrt{l_{\text{slide}}^2 + l_{\text{skip}}^2}$). We note that l_{1D} is what can be determined experimentally as the span of a lateral excursion, which is not equal to the variable l_{slide} - even when the protein only performs sliding. Namely, as we have defined l_{slide} to be the rms between consecutive skips, this quantity becomes much greater than l_{1D} if on average less than a skip occurs per search round ($n_{\text{slide}} \gg 1$ when $N_{\text{skip}} \ll 1$).

Taken together, the search time can be written as

$$\begin{aligned} T_{\text{search}} &= N_{\text{rnd}} T_{\text{rnd}} = N_{\text{rnd}} \left[T_{\text{slide}} + T_{\text{skip}} + \tau_{\text{3D}} \right] \\ &= L \times y \frac{(x\delta l)^2 \tau_{\text{slide}} + \left({}^{x}\!/\!{}_{y}\right)^2 \tau_{\text{skip}} + \tau_{\text{3D}}}{(1+y) x p_{\text{check}}(x) \delta l(y)} \end{aligned} \tag{S6.42}$$

In what follows, we shall first derive p_{check}, and proceed to show T_{search} has minima both for large scanning densities (sliding only) and low scanning densities (skip-n-slide).

Probability to interrogate all sites within a given section of sequence space

As discussed in the derivation leading up to **Equation ??**, after sufficient 'skip-and-slide cycles' the protein's motion is approximately described by a simple random walk with basic step length l_{sNs} and a probability ρ_{scan} to check all the bases within each segment per visit. Here, we derive an approximate equation for p_{check} for which we used Monte Carlo simulations to show it has the correct scaling with the model parameters (see main text and **Figure ??**) - thereby validating our analysis of the search time done below.

Let the protein bind to the DNA at segment 1 and leave it at $\hat{l}_{\text{1D}} = {}^{l_{\text{1D}}}\!/\!{}_{l_{\text{sNs}}}$. Towards calculating the probability to check all sites along its path at least once, we first pick a segment $\hat{l}$ between start- and endpoints and determine the probability to visit/interrogate all sites in this segment at least once prior to making it to segment $\hat{l}_{\text{1D}}$ for the first time (**Figure ??**A). Assuming the protein does not visit any other segments outside the interval $[1, \hat{l}_{\text{1D}}]$, the probability to reach $\hat{l}_{\text{1D}}$ after having checked the sites within $\hat{l}$ equals the probability of making it from $\hat{l}$ to $\hat{l}_{\text{1D}}$,

$$P(1 \to \hat{l}_{\text{1D}} | \text{check } \hat{l}) = P(1 \to \hat{l}) \times P(\hat{l} \to \hat{l}_{\text{1D}} | \text{check } \hat{l}) = P(\hat{l} \to \hat{l}_{\text{1D}} | \text{check } \hat{l}), \tag{S6.43}$$

as the protein will always return from the first segment to the intermediate (with or without checking sites in between) ($P(1 \to \hat{l}) = 1$). The probability of making it from 1 to $\hat{l}_{1d}$ *without* checking the intermediate site equals (**Figure ??**A)

$$\begin{aligned} P(\hat{l} \to \hat{l}_{\text{1D}} | \text{no check } \hat{l}) &= \frac{1}{2}(1-\rho_{\text{scan}}) P_{\text{no shuttle}}(\hat{l}_{\text{1D}} - \hat{l}) \\ &\times \sum_{m=0}^{\infty} \left(\frac{1}{2}(1-\rho_{\text{scan}}) \left[P_{\text{no shuttle}}(\hat{l}) + P_{\text{shuttle}}(\hat{l}) + P_{\text{no shuttle}}(\hat{l}_{\text{1D}} - \hat{l}) \right] \right)^m \\ &= \frac{1}{1 + \frac{2\rho_{\text{scan}}(\hat{l}_{\text{1D}} - \hat{l})}{1-\rho_{\text{scan}}}}, \end{aligned} \tag{S6.44}$$

with $P_{\text{shuttle}}(d)$ given by **Equation ??**. The common term in **Equation ??** represents the path that leads directly from segment $\hat{l}$ to the final one at $\hat{l}_{1D}$ without having checked the intermediate site. The first set of terms within the sum are all paths that attempt to reach segment 1, but do not make it (**Figure ??**A). The middle terms within the sum count all paths that do make it to the first segment and return with unit probability. The final term within the sum represents all paths that attempt to walk to the final segment, but do not make it across. From this we derive

$$P(\hat{l} \to \hat{l}_{\text{1D}}|\text{check } \hat{l}) = 1 - P(\hat{l} \to \hat{l}_{\text{1D}}|\text{no check } \hat{l}) = \frac{(\hat{l}_{\text{1D}} - \hat{l})}{(\hat{l}_{\text{1D}} - \hat{l}) + \frac{1-\rho_{\text{scan}}}{2\rho_{\text{scan}}}} \tag{S6.45}$$

As this holds for any segment within $[1, \hat{l}_{\text{1D}}]$, we get the probability of interrogating all sites/segments by averaging over all positions of $\hat{l}$,

$$\begin{aligned} p_{\text{check}}(\rho_{\text{scan}}, l_{\text{1D}}) &\approx \frac{1}{\hat{l}_{\text{1D}}} \int_0^{\hat{l}_{\text{1D}}} P(\hat{l} \to \hat{l}_{\text{1D}}|\text{check } \hat{l}) \mathrm{d}\hat{l} \\ &= 1 - \frac{1 - \rho_{\text{scan}}}{2\rho_{\text{scan}} l_{\text{1D}}/l_{\text{sNs}}} \log\left[1 + \frac{2\rho_{\text{scan}} l_{\text{1D}}/l_{\text{sNs}}}{1 - \rho_{\text{scan}}}\right], \end{aligned} \tag{S6.46}$$

for which we assumed large enough distances $\hat{l}_{\text{1D}}$ such that $\frac{1}{\hat{l}_{\text{1D}}} \sum_{\hat{l}=1}^{\hat{l}_{\text{1D}}} P(\hat{l} \to \hat{l}_{\text{1D}}|\text{check }) \approx \frac{1}{\hat{l}_{\text{1D}}} \int_1^{\hat{l}_{\text{1D}}} P(\hat{l} \to \hat{l}_{\text{1D}}|\text{check } \hat{l}) \mathrm{d}\hat{l} \approx \frac{1}{\hat{l}_{\text{1D}}} \int_0^{\hat{l}_{\text{1D}}} P(\hat{l} \to \hat{l}_{\text{1D}}|\text{check } \hat{l}) \mathrm{d}\hat{l}$.

We can rewrite **Equation ??** using $x = \frac{l_{\text{1D}}}{l_{\text{sNs}}} \frac{\rho_{\text{scan}}}{1-\rho_{\text{scan}}}$ (which is equal to **Equation ??**, by virtue of **Equation ??**),

$$p_{\text{check}}(x) = 1 - \frac{\log(1 + 2x)}{2x} \approx \begin{cases} x - \frac{4}{3}x^2 & x \ll 1 \\ 1 & x \gg 1 \end{cases} \tag{S6.47}$$

Conditions for optimal search time

We now proceed to find the optima of **Equation ??** in terms of x and y. Its derivative with respect to x equals

$$\partial_x \log T_{\text{search}} = \frac{2}{x} \frac{\tau_{\text{1D}}}{\tau_{\text{1D}} + \tau_{\text{3D}}} - \frac{1}{x} - \partial_x \log p_{\text{check}} \tag{S6.48}$$

Setting it equal to zero results in the following condition

$$\frac{2\tau_{\text{1D}}}{\tau_{\text{1D}} + \tau_{\text{3D}}} = 1 + x\partial_x \log p_{\text{check}} \tag{S6.49}$$

Similarly, setting $\partial_y \log T_{\text{search}}$ equal to zero results in

$$2\left[\frac{y}{1 + 2y} \frac{T_{\text{slide}}}{T_{\text{rnd}}} + \frac{T_{\text{skip}}}{T_{\text{rnd}}}\right] = 1 + \frac{y}{1 + 2y} - \frac{y}{1 + y} \tag{S6.50}$$

In what follows it is our goal to prove the existence of (at least) two minima - sets of coordinates in $\{N_{\text{slide}}, N_{\text{skip}}\}$-space, or equivalently $\{x, y\}$-space, that simultaneously satisfy **Equations ??** and **??**.

high scanning densities: sliding-only optimum

Here, we seek a local minimum of **Equation ??** - - satisfying both **Equations ??** and **??** - in the 'densely scanned' regime ($\rho_{\text{scan}} \gg 0.5$). For sufficiently large scanning densities, $y \gg 1$, for which **Equations ??** and **??** make the second term on the left hand side of **Equation ??** vanish, and we are left with

$$T_{\text{slide}} = \frac{1}{2} T_{\text{rnd}} \tag{S6.51}$$

If we additionally assume (close to) no skipping takes place ($N_{\text{skip}} \rightarrow 0$), or $y \gg x$ (**Equation ??**), this condition simplifies further to

$$\tau_{1\text{D}} = \tau_{3\text{D}} \tag{S6.52}$$

We see that at (close to) unit scanning density it is most beneficial to spend half of the time searching laterally along the substrate and the other half using excursions through solutions to reach distant sites. This result was obtained by Slutsky and Mirny [Slutsky and Mirny, Biophysical Journal 2004], whose model does not allow for skips to take place. Hence, our model coincides with theirs when shutting down skipping. Using **Equation ??** in **??** yields

$$x \partial_x p_{\text{check}} = 0 \tag{S6.53}$$

As this equation is satisfied both for $x \gg 1$ (**Equation ??**), and for $x = 0$ (using the $x \ll 1$ case in **Equation ??**), we identify the sliding-only case,

$$N_{\text{skip}}^{\text{sliding}} \rightarrow 0, \;\; N_{\text{slide}}^{\text{sliding}} = \frac{\tau_{3\text{D}}}{\tau_{\text{slide}}}, \;\; l_{\text{slide}}^{\text{sliding}} \gg l_{\text{skip}}^{\text{sliding}}, \; \rho_{\text{scan}}^{\text{sliding}} \rightarrow 1, \tag{S6.54}$$

as a (local) optimal search strategy. Recognizing that $l_{1\text{D}} = \sqrt{N_{\text{slide}}}$ for $N_{\text{skip}} = 0$ (**Equation ??**), and using **Equations ??**, **??** and **??** results in a search time (**Equation ??**) at the "sliding-only" optimum of

$$T_{\text{search}}^{\text{sliding}} = 2L\sqrt{\tau_{\text{slide}} \tau_{3\text{D}}} \tag{S6.55}$$

Hence, the search time can be minimized by eliminating skips altogether and adopting a scanning density of 1 ($l_{\text{slide}} \gg l_{\text{skip}}$).

low scanning densities: skipping-and-sliding optimum

Next, we seek to find an optimal search strategy that involves (frequent) skips. Returning to the y-derivative shown in **Equation ??**, we now explore the opposite limit of low scanning densities ($\rho_{\text{scan}} \ll 0.5$, $l_{\text{slide}} \ll l_{\text{skip}}$), $y \ll 1$, for which

$$T_{\text{skip}} = \frac{1}{2} T_{\text{rnd}} \tag{S6.56}$$

We see that at low scanning densities, it is most beneficial for the protein to spent half of its time interrogating sites following skips. Before proceeding, we introduce

$$\tau_{\text{slow}} = \tau_{\text{slide}} l_{\text{skip}}^2 \tag{S6.57}$$

as the time required to travel a full skipping length purely through sliding. That is, $\tau_{\text{skip}}/\tau_{\text{slow}} < 1$ indicates, after having taken into account the temporal cost of performing the skip, it remains beneficial to skip instead of just using sliding to reach the same region of the DNA/RNA. Having defined this variable, **Equation ??** results in

$$(y^{\text{sNs}})^2 = \underbrace{\frac{\tau_{\text{skip}}}{\tau_{\text{slow}}}}_{y_0^2} \frac{(x^{\text{sNs}})^2}{(x^{\text{sNs}})^2 + \underbrace{\frac{\tau_{\text{3D}}}{\tau_{\text{slow}}}}_{x_0^2}} = y_0^2 \frac{\left(x^{\text{sNs}}/x_0\right)^2}{1 + \left(x^{\text{sNs}}/x_0\right)^2}, \tag{S6.58}$$

where we have introduced x_0 and y_0 for notational convenience. Using this y-coordinate reduces **Equation ??** into a condition for the x-coordinate only

$$x\partial_x \log p_{\text{check}}|_{x=x^{\text{sNs}}} = \frac{\left(x^{\text{sNs}}/x_0\right)^2}{1 + \left(x^{\text{sNs}}/x_0\right)^2} \tag{S6.59}$$

Both sides of **Equation ??** are monotonic functions in x (**Figure ??**B). Hence, there is an optimal $T^{\text{sNs}}_{\text{search}}$ at $\{x^{\text{sNs}}, y^{\text{sNs}}\}$ corresponding to small scanning densities ($\rho_{\text{scan}} < 0.5$).

To obtain the corresponding value of the search time ($T^{\text{sNs}}_{\text{search}}$), we proceed to solve **Equation ??**. Although we are unable to solve **Equation ??** for general x, we can however obtain an approximate solution by assuming $x \ll 1$, for which (using **Equation ??** to simplify the left hand side of **Equation ??**)

$$\frac{(x^{\text{sNs}})^3}{2x_0^2} + x^{\text{sNs}} = \frac{3}{8} \tag{S6.60}$$

If we further assume $2x_0^2 \ll 1$, or equivalently, $\tau_{\text{3D}} \ll \tau_{\text{slow}}$,

$$x^{\text{sNs}} \approx \left(\frac{3}{4}\right)^{1/3} x_0^{2/3} \approx x_0^{2/3} = \left(\frac{\tau_{\text{3D}}}{\tau_{\text{slow}}}\right)^{1/3} \tag{S6.61}$$

To demonstrate the validity of this assumption we compared the numerical solution to **Equation ??** to the above approximation thereof (**Equation ??**). **Figure ??**C shows these to differ less than a factor 3 over a range in $\tau_{\text{3D}}/\tau_{\text{slow}}$ that spans 20 orders of magnitude. We therefore deem **Equation ??** to be valid also outside the $\tau_{\text{3D}}/\tau_{\text{slow}} \ll 1$ taken to obtain it initially (further allowing us to ignore the factor of $(3/4)^{1/3} \approx 0.91$). Using the x-coordinate, we obtain the following y-coordinate (**Equation ??**)

$$y^{\text{sNs}} = y_0 \sqrt{\frac{1}{1 + x_0^{2/3}}} \tag{S6.62}$$

Next, using that $\delta l \approx l_{\text{skip}}$ for $y \ll 1$ (the limit already taken), we find the following number of skipping and sliding steps taken in every search round (**Equations ??** and **??**)

$$N^{\text{sNs}}_{\text{slide}} = x_0^{4/3} l^2_{\text{skip}} = \frac{\tau_{\text{3D}}}{\tau_{\text{slide}}} \left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{1/3} \tag{S6.63}$$

$$N^{\text{sNs}}_{\text{skip}} = \frac{x_0^{4/3}}{y_0^2}\left(1 + x_0^{2/3}\right) = \frac{\tau_{\text{3D}}}{\tau_{\text{skip}}}\left(1 + \left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{1/3}\right) \tag{S6.64}$$

Combining **Equations ??** and **??** together with the skip-n-slide optimum set by **Equations ??**, **??**, **??** and **??** , results in a search time (**Equation ??**)

$$T_{\text{search}}^{\text{sNs}} = 2L\frac{\sqrt{\tau_{\text{skip}}\tau_{\text{3D}}}}{l_{\text{skip}}}\left(\frac{\sqrt{1+\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{1/3}}}{p_{\text{check}}(x^{\text{sNs}})}\right) \tag{S6.65}$$

In conclusion, the search time is minimized both at a maximum scanning density of 1 ($\rho_{\text{scan}}^{\text{sliding}} \approx 1$) - with a search time of $T_{\text{search}}^{\text{sliding}}$ (**Equation ??**) - and at a lower scanning density ($\rho_{\text{scan}}^{\text{sNs}} = \frac{1}{l_{\text{skip}}}\sqrt{N_{\text{slide}}^{\text{sNs}}/N_{\text{skip}}^{\text{sNs}}} < 0.5$) - with a search time $T_{\text{search}}^{\text{sNs}}$ (**Equation ??**).

Global Optimum

Having found two local optima, the more favorable search strategy is the one corresponding to the lowest search time. Hence, a combination of skipping and sliding is preferred (over just sliding) when $T_{\text{searh}}^{\text{sNs}} < T_{\text{search}}^{\text{sliding}}$. Using **Equations ??** and **??**

$$\frac{T_{\text{search}}^{\text{sNs}}}{T_{\text{search}}^{\text{sliding}}} = \sqrt{\frac{\tau_{\text{skip}}}{\tau_{\text{slow}}}}\left(\frac{\sqrt{1+\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{1/3}}}{p_{\text{check}}(x^{\text{sNs}})}\right) < 1 \tag{S6.66}$$

This can be rewritten as

$$\frac{\tau_{\text{skip}}}{\tau_{\text{slow}}} < \frac{p_{\text{check}}^{2}(x^{\text{sNs}})}{1+\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{1/3}} < 1 \tag{S6.67}$$

The second inequality ('less than 1') follows from noticing that $p_{\text{check}}(x) \leq 1$ for any x as it is a probability, and $\left(\frac{\tau_{\text{slow}}}{\tau_{\text{3D}}}\right)^{1/3} > 0$ as all τ's are positive, together making the middle identity always less than 1. As expected, $\frac{\tau_{\text{skip}}}{\tau_{\text{slow}}} < 1$, for skipping to be beneficial. However, **Equation ??** refines this statement and gives the exact boundary shown in the phase diagram of **Figure 3D**.

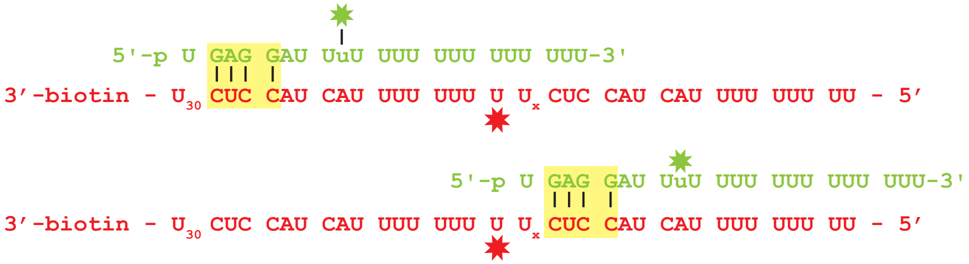
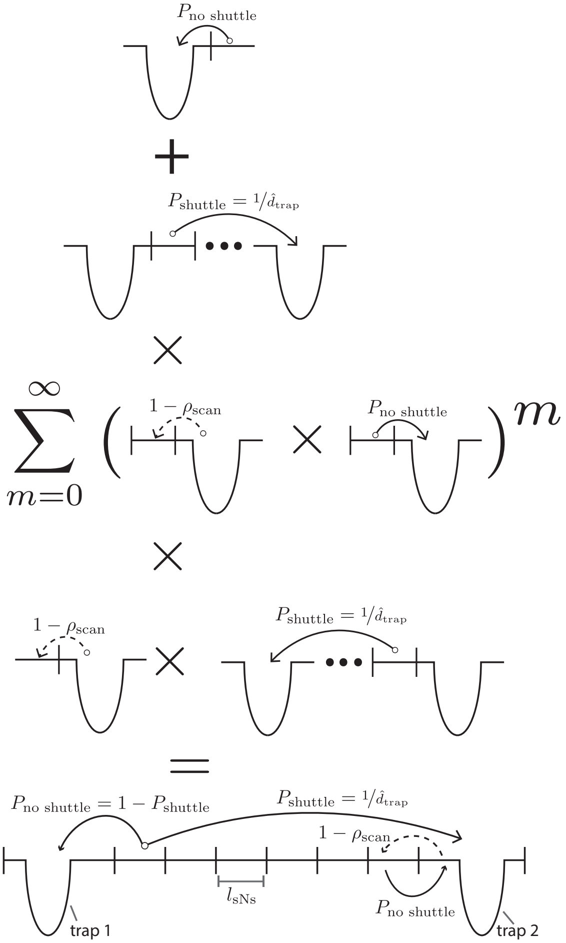

5'-p U GAG GAU UuU UUU UUU UUU UUU-3'
3'-biotin - U30 CUC CAU CAU UUU UUU U Ux CUC CAU CAU UUU UUU UU - 5'
5'-p U GAG GAU UuU UUU UUU UUU UUU-3'
3'-biotin - U30 CUC CAU CAU UUU UUU U Ux CUC CAU CAU UUU UUU UU - 5'

Figure S6.1: related to Figure ??. construct design hAgo2. ssRNA constructs (red) are passivated to the microscope slide using a 3'-biotin-streptavadin linkage. The two trapping sequences, 4 nt sequences that are complementary to the corresponding guide nucleotides (green), are highlighted in yellow. Top figure represents the 'high FRET' configuration, while the bottom figure displays Ago bound to the trap resulting in 'low FRET'. The distance between traps is varied by adding Uracil nucleotides (Ux reads: 'x times a U'). To embed the traps within the sequence, as opposed to them being the outermost sites, poly-U sequences flank both traps.



walk straight back into segment containing the first trap

“or...”

make it to segment *containing* the destination trap

“and..”

without interrogating the trap itself make it out of the segment.

walk back to the destination segment. repeat *m* times.

“and...”

Without interrogating the trap itself make it out of the segment and walk back into initial trap’s segment

Figure S6.2: related to Figure ??. path counting to derive scaling of shuttling time with distance. A graphical explanation of **Equation ??**. Subsequent figures will only show the equivalent of the bottom shown here.

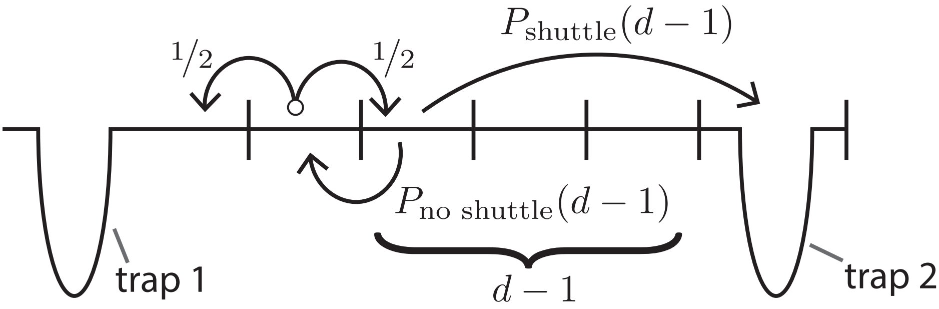


Figure S6.3: related to Figure ??. Shuttling time simple diffusion scales linearly with d_{trap}. Illustration of recursion relation dictating probability to shuttle P_{shuttle} (or get recaptured $P_{\text{no shuttle}}$) in terms of number of binding sites separating the two traps. Relates to **Equations ??** and **??**.

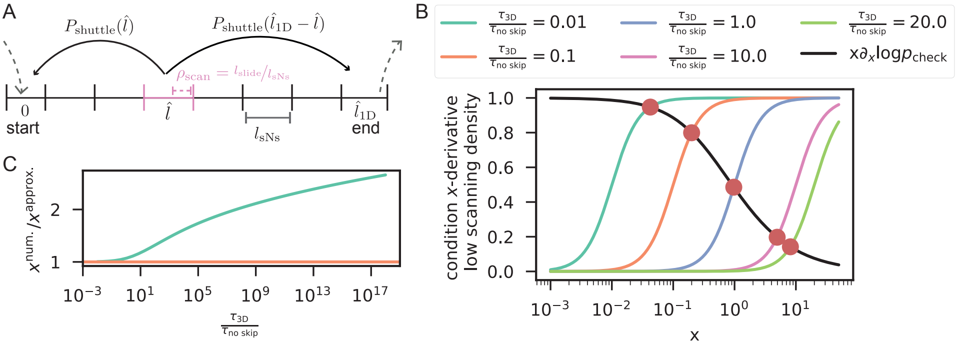


Figure S6.4: related to Figure ??.derivation of search time at given scanning density. (A) Illustration of paths (and corresponding probabilities) that lead the protein from segment 1 to $\hat{l}_{1D}$ (size l_{sNs}) without having interrogated all binding sites within segment $\hat{l}$. Relates to **Equation ??**. **(B)** At low scanning densities, the search time exhibits a unique minimum. Colored lines show right hand side of **Equation ??** for varying values of $\tau_{3D}/\tau_{\text{slow}}$ and black line shows the left hand side. Intersections (red dots) our found numerically and –together with **Equation ??** -indicate the location in $\{x, y\}$-space the skip-and-slide optimum can be found at (**Equation ??**). **(C)** Approximate location of skip-and-slide optimum (x-coördinate) from **Equation ??** versus numerical solution to **Equation ??**.

References

[] O. G. Berg, R. B. Winter, and P. H. von Hippel, *Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory.* Biochemistry **20**, 6929 (1981).

[] M. B. Elowitz, M. G. Surette, P.-e. Wolf, J. B. Stock, S. Leibler, M. B. Elowitz, M. G. Surette, P. E. Wolf, J. Stock, and S. Leibler, *Protein Mobility in the Cytoplasm of Escherichia coli,* Journal of Bacteriology **181**, 197 (1999).

[] S. E. Halford and J. F. Marko, *How do site-specific DNA-binding proteins find their targets?* Nucleic Acids Research **32**, 3040 (2004).

[] P. H. Vonhippel and O. G. Berg, *Facilitated Target Location in Biological-Systems,* Journal of Biological Chemistry **264**, 675 (1989).

[] A. D. Riggs, S. Bourgeois, and M. Cohn, *The lac represser-operator interaction. III. Kinetic studies,* Journal of Molecular Biology **53**, 401 (1970).

[] P. H. Richter and M. Eigen, *Diffusion controlled reaction rates in spheroidal geometry,* Biophysical Chemistry **2**, 255 (1974).

[] M. Slutsky and L. A. Mirny, *Kinetics of Protein-DNA Interaction : Facilitated Target Location in Sequence-Dependent Potential,* Biophysical Journal **87**, 4021 (2004).

[] P. C. Blainey, A. M. van Oijen, A. Banerjee, G. L. Verdine, and X. S. Xie, *A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA,* Proceedings of the National Academy of Sciences **103**, 5752 (2006).

[] I. Bonnet, A. Biebricher, P. L. Porté, C. Loverdo, O. Bénichou, R. Voituriez, C. Escudé, W. Wende, A. Pingoud, and P. Desbiolles, *Sliding and jumping of single EcoRV restriction enzymes on noncognate DNA,* Nucleic Acids Research **36**, 4118 (2008).

[] D. M. Gowers, G. G. Wilson, and S. E. Halford, *Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA,* Proceedings of the National Academy of Sciences **102**, 15883 (2005).

[] A. Graneli, C. C. Yeykal, R. B. Robertson, and E. C. Greene, *Long-distance lateral diffusion of human Rad51 on double-stranded DNA,* Proceedings of the National Academy of Sciences **103**, 1221 (2006).

[] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf, *The lac Repressor Displays Facilitated,* Science **336**, 1595 (2012).

[] J. H. Kim and R. G. Larson, *Single-molecule analysis of 1D diffusion and transcription elongation of T7 RNA polymerase along individual stretched DNA molecules,* Nucleic Acids Research **35**, 3848 (2007).

[] A. B. Kochaniak, S. Habuchi, J. J. Loparo, D. J. Chang, K. A. Cimprich, J. C. Walter, and A. M. van Oijen, *Proliferating cell nuclear antigen uses two distinct modes to move along DNA,* Journal of Biological Chemistry **284**, 17700 (2009).

[] H. R. Koh, M. A. Kidwell, J. Doudna, and S. Myong, *RNA scanning of a molecular machine with a built-in ruler,* Journal of the American Chemical Society **139**, 262 (2017).

[] J. S. Leith, A. Tafvizi, F. Huang, W. E. Uspal, P. S. Doyle, A. R. Fersht, L. A. Mirny, and A. M. van Oijen, *Sequence-dependent sliding kinetics of p53,* Proceedings of the National Academy of Sciences **109**, 16552 (2012).

[] D. Normanno, L. Boudarène, C. Dugast-Darzacq, J. Chen, C. Richter, F. Proux, O. Bénichou, R. Voituriez, X. Darzacq, and M. Dahan, *Probing the target search of DNA-binding proteins in mammalian cells using TetR as model searcher,* Nature Communications **6** (2015), 10.1038/ncomms8357.

[] K. Ragunathan, C. Liu, and T. Ha, *RecA filament sliding on DNA facilitates homology search,* eLife , 1 (2012).

[] N. P. Stanford, M. D. Szczelkun, J. F. Marko, and S. E. Halford, *One- and three-dimensional pathways for proteins to reach specific DNA sites.* The EMBO journal **19**, 6546 (2000).

[] A. Tafvizi, F. Huang, A. R. Fersht, L. A. Mirny, and A. M. van Oijen, *A single-molecule characterization of p53 search on DNA,* Proceedings of the National Academy of Sciences **108**, 563 (2011).

[] Y. M. Wang, R. H. Austin, and E. C. Cox, *Single molecule measurements of repressor protein 1D diffusion on DNA,* Physical Review Letters **97**, 1 (2006).

[] L. Zandarashvili, A. Esadze, D. Vuzman, C. A. Kemme, Y. Levy, and J. Iwahara, *Balancing between affinity and speed in target DNA search by zinc-finger proteins via modulation of dynamic conformational ensemble,* Proceedings of the National Academy of Sciences **112**, E5142 (2015).

[] J. Elf, G.-W. Li, and X. S. Xie, *Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell,* Science **316**, 1191 (2007).

[] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'Conner, D. W. Noble, and P. H. Von Hippel, *Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound Escherichia coli lac repressor in vivo,* Proceedings of the National Academy of Sciences **74**, 4228 (1977).

[] T. Hu and B. I. Shklovskii, *How a protein searches for its specific site on DNA: The role of intersegment transfer,* Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **76**, 1 (2007).

[] G.-w. Li, O. G. Berg, and J. Elf, *Effects of macromolecular crowding and DNA looping on gene regulation kinetics,* Nature Physics **5**, 294 (2009).

[] M. A. Lomholt, B. van den Broek, S.-M. J. Kalisch, G. J. L. Wuite, and R. Metzler, *Facilitated diffusion with DNA coiling,* Proceedings of the National Academy of Sciences **106**, 8204 (2009).

[] M. Sheinman and Y. Kafri, *The effects of intersegmental transfers on target location by proteins,* Physical Biology **016003** (2009), 10.1088/1478-3975/6/1/016003.

[] M. Klein, S. Chandradoss, M. Depken, and C. Joo, *Why Argonaute is needed to make microRNA target search fast and reliable,* Seminars in Cell and Developmental Biology **65**, 20 (2017).

[] B. van den Broek, M. A. Lomholt, S.-M. J. Kalisch, R. Metzler, and G. J. L. Wuite, *How DNA coiling enhances target localization by proteins,* Proceedings of the National Academy of Sciences **105**, 15738 (2008).

[] H. Chen, S. P. Meisburger, S. A. Pabit, J. L. Sutton, W. W. Webb, and L. Pollack, *Ionic strength-dependent persistence lengths of single-stranded RNA and DNA,* Proceedings of the National Academy of Sciences **109**, 799 (2012).

[] T. A. L. I. Azam, A. Iwata, A. Nishimura, S. Ueda, and A. Ishihama, *Growth Phase-Dependent Variation in Protein Composition of the,* Society **181**, 6361 (1999).

[] K. Brogaard, L. Xi, J. P. Wang, and J. Widom, *A map of nucleosome positions in yeast at base-pair resolution,* Nature **486**, 496 (2012).

[] V. Globyte, S. H. Kim, and C. Joo, *Single-Molecule View of Small RNA – Guided Target Search and Recognition,* Annual Review of Biophysics , 1 (2018).

[] H. Wang, M. La Russa, and L. S. Qi, *CRISPR/Cas9 in Genome Editing and Beyond,* Annual Review of Biochemistry **85**, 227 (2016).

[] S. D. Chandradoss, N. T. Schirle, M. Szczepaniak, I. J. Macrae, and C. Joo, *A Dynamic Search Process Underlies MicroRNA Targeting,* Cell **162**, 96 (2015).

[] V. Globyte, S. H. Lee, T. Bae, J. Kim, and C. Joo, *CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion,* The EMBO Journal **38**, e99466 (2019).

[] Y. Jeon, Y. H. Choi, Y. Jang, J. Yu, J. Goo, G. Lee, Y. K. Jeong, S. H. Lee, I. S. Kim, J. S. Kim, C. Jeong, S. Lee, and S. Bae, *Direct observation of DNA target searching and cleavage by CRISPR-Cas12a,* Nature Communications **9** (2018), 10.1038/s41467-018-05245-x, arXiv:15334406 .

[] A. A. Shvets and A. B. Kolomeisky, *Mechanism of Genome Interrogation: How CRISPR RNA-Guided Cas9 Proteins Locate Specific Targets on DNA,* Biophysical Journal **113**, 1416 (2017).

[] S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna, *DNA interrogation by the CRISPR RNA-guided endonuclease Cas9,* Nature **507**, 62 (2014), arXiv:NIHMS150003 .

[] J. W. Hegge, D. C. Swarts, S. D. Chandradoss, T. J. Cui, J. Kneppers, M. Jinek, C. Joo, and J. van der Oost, *DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute,* Nucleic Acids Research **47**, 5809 (2019).

[] T. J. Cui, M. Klein, J. W. Hegge, S. D. Chandradoss, J. van der Oost, M. Depken, and C. Joo, *Argonaute bypasses cellular obstacles without hindrance during target search,* bioRxiv (2019).

[] M. Bauer and R. Metzler, *Generalized facilitated diffusion model for DNA-binding proteins with search and recognition states,* Biophysical Journal **102**, 2321 (2012).

[] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj, *How a protein searches for its site on DNA : the mechanism of facilitated diffusion,* Journal of Physics A: Mathematical and Theoretical **42**, 1 (2009).

[] C. Joo and T. Ha, *Single-molecule FRET with total internal reflection microscopy,* Cold Spring Harbor Protocols **7**, 1223 (2012).

[] I. M. Silverman, F. Li, A. Alexander, L. Goff, C. Trapnell, J. L. Rinn, and B. D. Gregory, *RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome,* Genome Biology **15**, 1 (2014).

[] M. D. Biggin, *Animal Transcription Networks as Highly Connected, Quantitative Continua,* Developmental Cell **21**, 611 (2011).

[] A. Mahmutovic, O. G. Berg, and J. Elf, *What matters for lac repressor search in vivo - Sliding, hopping, intersegment transfer, crowding on DNA or recognition?* Nucleic Acids Research **43**, 3454 (2015).

[] Y. S. Dagdas, J. S. Chen, S. H. Sternberg, J. A. Doudna, and A. Yildiz, *A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9,* Science Advances **3**, 1 (2017).

[] S. H. Sternberg, B. Lafrance, M. Kaplan, and J. A. Doudna, *Conformational control of DNA target cleavage by CRISPR-Cas9,* Nature **527**, 110 (2015).

[] M. Yang, S. Peng, R. Sun, J. Lin, N. Wang, and C. Chen, *The Conformational Dynamics of Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET,* Cell Reports **22**, 372 (2018).

[] S. D. Chandradoss, A. C. Haagsma, Y. K. Lee, J.-H. Hwang, J.-M. Nam, and C. Joo, *Surface Passivation for Single-molecule Protein Studies,* Journal of Visualized Experiments , 1 (2014).

[] H. Pan, Y. Xia, M. Qin, Y. Cao, and W. Wang, *A simple procedure to improve the surface pas-*

sivation for single molecule fluorescence studies, Physical Biology **12** (2015), 10.1088/1478-3975/12/4/045006.

Summary

The past decade has witnessed a revolution in genome-engineering. Using CRISPR-Cas9 DNA sequences can be marked, detected and cleaved. Rewriting life's instructions in such a fashion paves the way towards numerous scientific, agricultural and medical applications. Without proper quantification of the associated risks we face the danger of applying treatments without knowing its consequences. Most notable concern lies in Cas9's specificity. Although Cas9 targets DNA complementary to any designed 20nt guide RNA, it notoriously also acts on non-fully matching sequences. This thesis describes work towards a physical understanding of how Cas9 and similar RNA/DNA guided systems locate and recognize their target. **Chapter** 1 introduces the reader to life's most important molecules (DNA, RNA and protein) as well as to the RNA guided CRISPR and Argonaute (Ago) systems. The chapter also provides an introduction to the main modeling techniques used in subsequent chapters.

In **Chapters** 2 and 3 we model the physics governing target selection. Our current understanding of binding and cleavage specificity is reflected in a set of rules of thumb used to design the 20nt target. **Chapter** 2 shows said rules are a direct consequence of having a unidirectional binding process, as assumed to be the case for both Cas9 and Argonaute. At the core of the presented model lies the free-energy landscape underlying the protein-guide-target interactions. **Chapter** 2 uses a simple landscape in which the addition of a matching base pair to the guide-target hybrid kinetically (as well as energetically) favors cleavage, while a mismatch makes rejection of the (off-)target more likely. With a single gain/penalty for every match/mismatch between guide and target we highlight the benefit of using a kinetic modeling approach. In **Chapter** 3, the parameterization is expanded to allow for position dependent (mis-)match biases, which are extracted from a series of high-throughput experimental datasets to elucidate in more detail the free-energy landscape of spCas9-sgRNA-DNA. The determined landscape directly explains what off-target sites are expected to lead to stable binding on timescales much shorter than cleavage, explaining the previously reported discrepancy between binding and cleavage specificities. Moreover, the free-energy landscape is consistent with single-molecule fluorescence experiments probing the conformational dynamics of Cas9 during target binding, thereby showing how Cas9's major conformational change couples to the hybrid-formation process. Finally, this chapter demonstrates how our kinetic model improves upon existing target prediction tools.

Chapters 4-6 describe a protein's search for a single target site embedded within the genome. **Chapter** 4 reviews literature describing how target searching proteins use a combination of three-dimensional diffusion through solution with (effective) one-dimensional diffusion along the contour of the DNA. Furthermore, using the human Argonaute 2 protein as an example, **Chapter** 4 hypothesizes how coupling structural changes to hybrid formation, as we also show for spCas9 in **Chapter** 3, can balance search time and specificity. **Chapters** 5 and 6 present a collaboration with experimentalist from the lab of Chirlmin Joo. First,

Chapter 5 shows a prokaryotic Argonaute can bypass other DNA bound proteins when laterally scanning the DNA. As a model in which Ago is forced to interrogate all binding sites during a lateral excursion cannot account for the measured diffusion rates, bases must have been skipped when moving past the protein blockades. Next **Chapter** 6 describes a model allowing for such base skipping, resulting in only a fraction of the DNA enclosed within a lateral excursion being interrogated. Additional single-molecule experiments show that also human Ago uses such base-skipping. Although both Ago only interrogate all DNA after many repeated rounds of lateral diffusion, we show such a mechanism helps to speed up the search for the cognate target.

Samenvatting

In het afgelopen decennium heeft zich een revolutie in de genoombewerkingstechnologie voltrokken. Gebruikmakend van CRISPR-Cas9 kan DNA worden opgespoord en geknipt, waarmee verschillende wetenschappelijke, agrarische en medische toepassingen een stap dichterbij zijn. Als de mogelijke risico's van deze krachtige techniek niet worden gekwantificeerd bestaat de angst dat medische behandelingen plaatsvinden zonder dat alle mogelijke gevolgen bekend zijn. Het grootste risico zit in de specificiteit van Cas9. In principe wordt Cas9 geprogrammeerd om DNA te knippen met een sequentie van 20nt complementair aan een aan de proteïne meegegeven 'gids' RNA. Helaas knipt Cas9 ook DNA-sequenties die niet compleet complementair zijn aan de gids. Het werk omschreven in dit proefschrift draagt bij aan fysisch inzicht in de manier waarop Cas9, en soortgelijke RNA/DNA geprogrammeerde systemen, hun doelwit DNA vinden en herkennen. **Hoofdstuk** 1 maakt de lezer bekent met de meest belangrijke biomoleculen (DNA, RNA en eiwitten) en de CRISPR en Argonaute (Ago) systemen die in dit proefschrift uitvoerig bestudeerd zijn. Tevens bevat dit hoofdstuk een introductie tot de wiskundige technieken die gebruikt zijn voor het opstellen van de modellen verderop in het proefschrift.

Hoofdstukken 2 en 3 presenteren een fysisch model dat omschrijft hoe Cas9 en Ago hun doelwit herkennen. Ons huidig begrip van de specificiteit van dit soort systemen kan worden samengevat met een aantal vuistregels die in acht worden genomen bij het ontwerpen van het gis RNA. **Hoofdstuk** 2 laat zien dat deze regels een direct gevolg zijn van een bindingsproces dat aan een kant van de gids begint, zoals aangenomen wordt het geval te zijn voor zowel Cas9 als Ago. In het model staat het vrije-energielandschap dat interacties tussen gids RNA, doelwit DNA en het eiwit omschrijft centraal. In **Hoofdstuk** 2 wordt er een simpel landschap gebruikt waarin de toevoeging van een complementair basepaar aan de gids-doelwit hybride een kinetisch (alsmede een energetisch) voordeel oplevert. De toevoeging van een non-complementair basepaar verhoogt de waarschijnlijkheid dat de proteïne ontbindt. Dit simpele landschap, met een enkel voordeel/nadeel voor een correct/incorrect basepaar belicht het voordeel van het gebruik van een kinetisch model.

In **Hoofdstuk** 3 wordt de parameterizatie uitgebreid. Gebruikmakend van experimentele datasets, worden de positieafhankelijke voordelen/nadelen voor correcte/incorrecte baseparen geëxtraheerd, waaruit een meer gedetailleerd vrije-energielandschap van spCas9-sgRNA-DNA volgt. Dit landschap verklaard hoe bij sommige non-complementaire DNA doelwitten Cas9-gids stabiel bindt, lang voordat er geknipt wordt. Hiermee geven wij een verklaring voor het verschil in de schijnbare specificiteit van het binden van inactief Cas9 en het knippen van actief/inactief Cas9. Het vrije-energielandschap is tevens consistent met fluorescentie experimenten die de eiwitconformatie van Cas9 tijdens het bindingsproces bestuderen, waardoor het gepresenteerde landschap direct laat zien hoe de grootste verandering van conformatie koppelt aan het bindingsproces tussen gids en DNA. Ten slotte laat dit hoofdstuk zien hoe ons kinetisch model een verbetering over bestaande modellen biedt.

Hoofdstukken 4-6 beschrijven de zoektocht van een eiwit naar een enkel correct doelwit binnen een vele malen groter genoom. **Hoofdstuk** 4 biedt een beschouwing van de literatuur waarin beschreven wordt dat eiwitten hun doelwit vinden doormiddel van een combinatie van driedimensionale diffusie door oplossing en (effectieve) eendimensionale diffusie langs de contour van het DNA. **Hoofdstuk** 4 brengt tevens het idee naar voren dat een koppeling tussen eiwitconformatie en het bindingsproces gebruikt kan worden om het doelwit zowel snel als specifiek te herkennen. Het hoofdstuk gebruikt het menselijke Argonaute 2 als zo een systeem met een dergelijke koppeling, net als **Hoofdstuk** 3 eenzelfde soort koppeling suggereert voor Cas9.

Hoofdstukken 5 en 6 zijn uitgevoerd in samenwerking met experimentalisten uit het lab van Chirlmin Joo. **Hoofdstuk** 5 laat zien dat een prokaryotische Argonaute andere aan het DNA gebonden eiwitten kan omzeilen. De experimenten laten zien dat Ago sneller langs het DNA diffundeert, dan een model waarin Ago iedere sequentie langs het DNA vergelijkt met zijn gids voorspelt. Hieruit concluderen we dat Ago sequenties langs het DNA overslaat om zo obstakels langs het DNA te vermijden. **Hoofdstuk** 6 bouwt hierop voort door een model op te stellen waarin Ago ook sequenties kan overslaan, waardoor slechts een fractie van het DNA waarlangs diffundeert wordt daadwerkelijk vergeleken wordt met de gids. Aanvullende enkel-molecuul experimenten laten zien dat ook het menselijke Ago DNA sequenties overslaat. Ondanks dat vele rondes van laterale diffusie nodig zijn alvorens Ago alle mogelijke DNA sequenties kan hebben vergelijken met de gids, laat dit laatste hoofdstuk zien dat dit eigenlijk helpt om het correcte doelwit sneller te vinden.

Acknowledgements

A mixture of gratefulness, nostalgia, and happiness takes control over me as I now attempt to write that part of this thesis that, let's be honest, will probably be the most read part. The last couple of years have brought me in contact with so many wonderful people and made me experience things I never in a million years thought I would.

Martin, thanking you for being a good supervisor would be a strong understatement. Before starting my PhD, I got to experience working for you a little bit. I immediately noticed you were able to get more out of me than I thought I had in me. This continued during the PhD. Other than teaching me most of the technical skills, you were a constant motivator, making me comfortable with voicing my ideas. It was fun to see the group grow as the project advanced. Thank you for the many (extremely) long meetings at the whiteboard ("the lab"). Perhaps the quality that really makes you a great supervisor is how much you genuinely care about your group members. Best of luck to you, Olya, and Markus.

Chirlmin, at moments you acted as my second supervisor. As a collaborator, you quickly showed your great expertise. Perhaps a more underrated quality are your excellent people's skills. You are a kind person, who also genuinely cares about his lab members, which is probably why you run an overall happy group of people. To my great surprise, at several occasions you cared to look out for me, even though I was 'just the collaborator'. Working with you and your lab members has been a pleasure thus far, and I am happy to continue doing such.

A huge part of the work done could not have been done without **Behrouz**, my 'partner in crime'. I got to know you as an extremely smart scientist. More importantly, you manage to combine this with a great sense of humor and the friendliest of smiles with which you start off each working day. Together this makes for one hell of an awesome teammate. When starting the project, we used to worry whether anything would ever work at all. Slowly, but surely, our conversations on our way out became more optimistic. I enjoyed going through the whole process together. Thank you for all the times you made me feel more positive about our work, for the countless of hours spend discussing CRISPR, and for always being in for a good laugh or chat. Best of luck to you, your mom, 'the turtle(s)', 'the donkey', 'the horse', Naazi, Baarfi and all the other animals. Khodoffess!!

I would like to thank my committee members **Ilya Finkelstein**, **David Rueda**, **Sander Tans**, **Pieter Rein ten Wolde**, **Helmut Schiessel** and **Marileen Dogterom** for spending their time reading this thesis, making there way to Delft, and for taking part in the defense.

Luckily I will not have to go through this defense all by myself and have a pair of terrific people as my paranymphs. **Nicole**, your great sense of humor and very kind personality made this final year and a half so much more enjoyable for me. Moreover, you probably more than anyone showed me how to live by ‘work hard, play hard’. Your dedication to such a difficult project really inspired me. I am 100% confident you will succeed in finishing your PhD and in whatever comes after. Thanks for the partying, turning one of my recommended Spotify playlists into Latin songs only, the music making and always being in for a good chat or laugh.

Thijs, we met the very first day of our Bsc studies. Waiting in line for our books, we bonded over how difficult we thought that first lecture was and remained friends ever since. It is therefore only suiting that we managed to make things come full circle by having our defenses one day apart. Good luck! Don’t mess up! Most importantly, do not (!), I repeat, do not(!), ask your paranymph to answer the question for you! Ow, and do not mess up. No pressure... Although, I think not much could pressure you even if one tried. You see, this is probably why working with you as a collaborator was so easy and enjoyable. Normally I would maybe not advice to be collaborators with your friends. Unless, that friend is Thijs. Thijs maneuvered along this balancing chord as if it was the widest boulevard in the world. When the project became tough or developing the theory took ‘a bit’ longer, you always managed to put things into perspective. Keeping your cool, you’d maybe complaint alongside me for 5 minutes after which you would simply give me the look that says ‘ah well, guess that has to happen’. This ‘stick-to-itiveness’ really helped getting the target search chapters done. Thank you.

Next, I’d like to express my gratitude to **The JooC lab**. Being part of a small group, you guys took me in and made your group feel like ‘a home away from home’. Thank you for all the time spent after office hours together, visiting me in the hospital and traveling all around the globe for conferences. **Iasonas**, my theory mate, happy to see the project will be in safe hands, and happy to work with you. Good luck balancing Delft and Athens! **Viktorija**, thank you for discussing CRISPR related topics and for making sure I would always get all the Cas9 facts straight before asking advice. It was always comforting, and funny, to have a familiar face presenting together with me during conferences. **Mike**, the true embodiment of *’geen worden, maar daden’*. Your ‘go-getter’ mentality will surely help you succeed. **Laura**, meeting you for a second time in Delft was a great surprise. Thanks for your sense of humor. **Ivo**, let’s get this software to work! (Although you are doing pretty good without my help). **Carolien**, welcome and good luck embarking on this high-throughput journey. **Sunchul**, aka ‘doctor Kim’, aka Koen, you are truly one of a kind! **Adi** and **Sung Hyun**, welcome back! Great to see a goodbye doesn’t always need to last forever, even when working in science! **Ilya**, I wish you (more than) the best of health the building has to offer. **Margreet**, thanks for letting me(!) teach a lab practical. I enjoyed the group meetings with all of you thus far and am looking forward to also get to know **Adam**, **Alessia**, **Carlos**, **Cecilia**, **Nathalie**, **Pim** and **Roy** better.
Thanks to the former JooCs. **Luuk**, thanks for helping me with CRISPR related stuff as well as being a cool conference buddy. **Stanley**, ‘the manley’, I very much enjoyed working together with you on the review chapter. **Mo**, thanks for all the nice talks. I am sure you will make a PI people are happy to work for.

Making music, especially together with others, is one of the most fun things to do. I am extremely thankful for all of those with who I've gotten the chance to do such. **Fede** and **Helena**, thank you so much for 'getting me back on the horse'. **Fede**, your such a kind person, which you showed by visiting me in the hospital and asking me if I would like to have my melodica with me. At the time, I did not have the long capacity to walk more than two steps, but the gesture shows your kind nature. Musically, I like how your emotional character is really translated into your singing and guitar playing. Good luck rapping things up, it was not an easy road for you! **Helena**, Shomi , woooowoop! Thanks for giving me the confidence to play just with you and perform the sometimes beautiful, sometimes silly songs. Thanks also for making me laugh during rehearsals. Again a great example of 'work hard, play hard', completing a not at all easy project along the way. Enjoy Tel Aviv! **Ale**, grande! Simply put: You Rock! **Sandro**, you once again proved to me that I tend to get along with bass players :P. **Fillip**, sorry for enduring us amateurs. **Sam**, last time we played I completely 'forgot' to play myself as I just got lost listening to your cello. Good luck fixing those tweezers. **JK**, **Louis**, **Anthony** 'the voice' Birny, **Sabina** (with an equally beautiful voice) thank you all for the creative outlet after work. Also thanks to **Hirad**, **Yiteng**, **Da**, and all other regular contributors to the BN music making festivities.

Over the course of the past 4.5 years I had the pleasure of being helped by great Bsc/Msc students. **Diego Gonzalez-Arroyo**, invaluable help during the very beginnings of the PhD project. The Maths you helped me with later turned out to be the solution to chapters 5 and 6 as well. **Michiel** , **Luuk van Duuren**, thank you for devotion to my 'Friday afternoon' ideas. **Christian**, such an organized and dedicated student that thought Behrouz and me about PR-curves. Thank you for the magic acts! **Koen**, thank you for your great help setting up the fitting code we now so heavily use. **Stijn**, again an excellent student. Thanks for helping during the final stages. **Diewertje**, I admired your devotion to a difficult project for a Bsc student. Thank you for teaching me to have a life outside the lab, also during the busy final stages of a PhD. **Sonny**, great to have you back and healthy. Let's make this work!

BN really is a pleasant working environment. I would like to thank **Marileen** for her efforts that created this melting pot of great people that output great science. In large part this is made possible by BN's silent forces, amongst who: **Tahnee**, **Amanda**, **Tracey**, **Jolijn**, **Dijana**,**Marije** and **Chantal** who always think alongside us and help swiftly.
Special thanks to both **Christophe** and **Cees**, for who without their help I would probably not even have been in the position to start my PhD. Thanks to all the other PIs **Bertus**, **Greg**, **Marie-Eve**, **Stan**, **Nynke**, **Timon**, **Arjen**, **Gijsje**, **Liedewij**, **Dimphna**, and **Hyun**. If not for simply asking me from time to time how things are going, well... for hiring the cool people!

So many cool people at BN! **Daniel V.**, hands-down the most clever person I got to work with. You taught me the valuable lesson of always staying positive. If not for your help beforehand, I probably would not have been ready to start my PhD. **Mathia**, **Lisa**, **Carsten** and **Benjamin**, I have kept the card, made me laugh. **Mathia**, such fortitude, amazing! Thank you for making the JooCs adopt others. **Richard**, it was really cool to see you and Behrouz succeed after so many years working on a project. Good luck with your next move, hopefully you get the job you are seeking. **Seb**, **Jochem**, **Cristobal** and **Becca**, the CRISPR guru's. It was real fun meeting some of you guys in New York and Israel prior to your move from Wageningen. It was a pleasure being at conferences you. Becca, thanks for your efforts to boost BN's social life. **Reza**, **Fayezeh**, **Afshin**, thank you on Behrouz' behalf for the 'tea time'. This really brought him back refreshed and full of new ideas. **Afshin**, thanks for all the coffee/tea breaks and the always pleasant chats. **Cátia**, **Kasper**, thanks for making our office a happy place. **Diego**, you too are responsible for the change in my Spotify list. **George**, thanks for answering my technical questions. My thanks to the so many others whome I've had the pleasure meating, talking and learning about their science: **Mehran**, **Ferhat**, **Franklin**, **David**, **Johannes**, **Duco**, **Elisa**, **Mitasha**, **Siddarth**, **Eugene**, **Biswajit**, **Daniel Xi**, **Sonja**, **Michel**, **Wayne**, **Alberto**, **Kayley**, **Anuj**, **Sumit**, **Essengül**. .

Michiel, **Mathijs**, **Adriaan**, **Nick**, thanks for the awesome trips to Tel Aviv, Helsinki and Limburg, even though the latter might have turned two of us into vegetarians :P. Good luck with completing your PhD's/ your carriers.

Aram!, 'my brotha from a Kurdish motha', good luck with your studies, **Beau**, sorry (not) for all the medical costs :P, **Chloé**, good luck becoming a surgeon, **Paula**, **Wria**. Cannot imagine the past few years without you guys. Thank you for the visits, walking, the crazy situations you (read: *Aram*) got me into, and just being good friends.

Mom and **Dad**. Thank you for always believing in me, even when I stopped doing so myself.

Curriculum Vitæ

Misha KLEIN

27-09-1991	Born in Chicago, USA.
2003–2009	Secondary Education, St. Laurens college, Rotterdam, The Netherlands
2009–2012	Bsc. Applied Physics Delft University of Technology, The Netherlands
2012-2015	Msc. Applied Physics Delft University of Technology, The Netherlands
2015-2019	PhD. Departement of BioNanoScience Kavli Institute of Nanoscience, Delft University of Technology, The Netherlands ***Thesis:*** Right Place, Right Time: Modeling the search time and specificity of Cas9 and Argonaute ***Copromotor:*** Dr. S. M. Depken ***Promotor:*** Dr. C. Joo

List of Publications

1. **Klein, M.**,Eslami-Mosslam, B., van der Smagt, S., Depken, M. *Mechanistic modeling explains dCas9 binding and Cas9 cleavage dynamics*, manuscript in preparation.

2. **Klein, M.**, Cui, T.J., Joo, C., Depken, M. *Optimal DNA/RNA target search using frequent skip-n-slides.* manuscript in preparation.

3. Cui, T.J., **Klein, M.**, Hegge, J., Chandradoss, S.D., van der Oost, J., Depken, M., Joo., C. *Argonaute bypasses cellular obstacles without hindrance during target search*, Nature Communications (2019).

4. **Klein, M.**, Eslami-Mosslam, B., Gonzalez-Arroyo, D., Depken, M.. *Hybridization Kinetics Explains CRISPR-Cas Off-targeting Rules.* Cell Reports (2018).

5. **Klein, M.**, Chandradoss, S.D., Depken M., Joo C. *Why Argonaute is needed to make microRNA target search fast and reliable*, Seminars in Cell and Developmental Biology (2017).

6. Künne, T., Kieper, S.N.,Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R. **Klein, M.**, Depken, M., Suarez-Diez, M.,Brouns, S.J.J. *Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation*, Molecular Cell (2016).

7. Nicoli, F., Verschueren, D. , **Klein, M.**, Dekker, C., Johnson, M.P. *DNA Translocations through Solid-State Plasmonic Nanopores*, Nanoletters (2014).