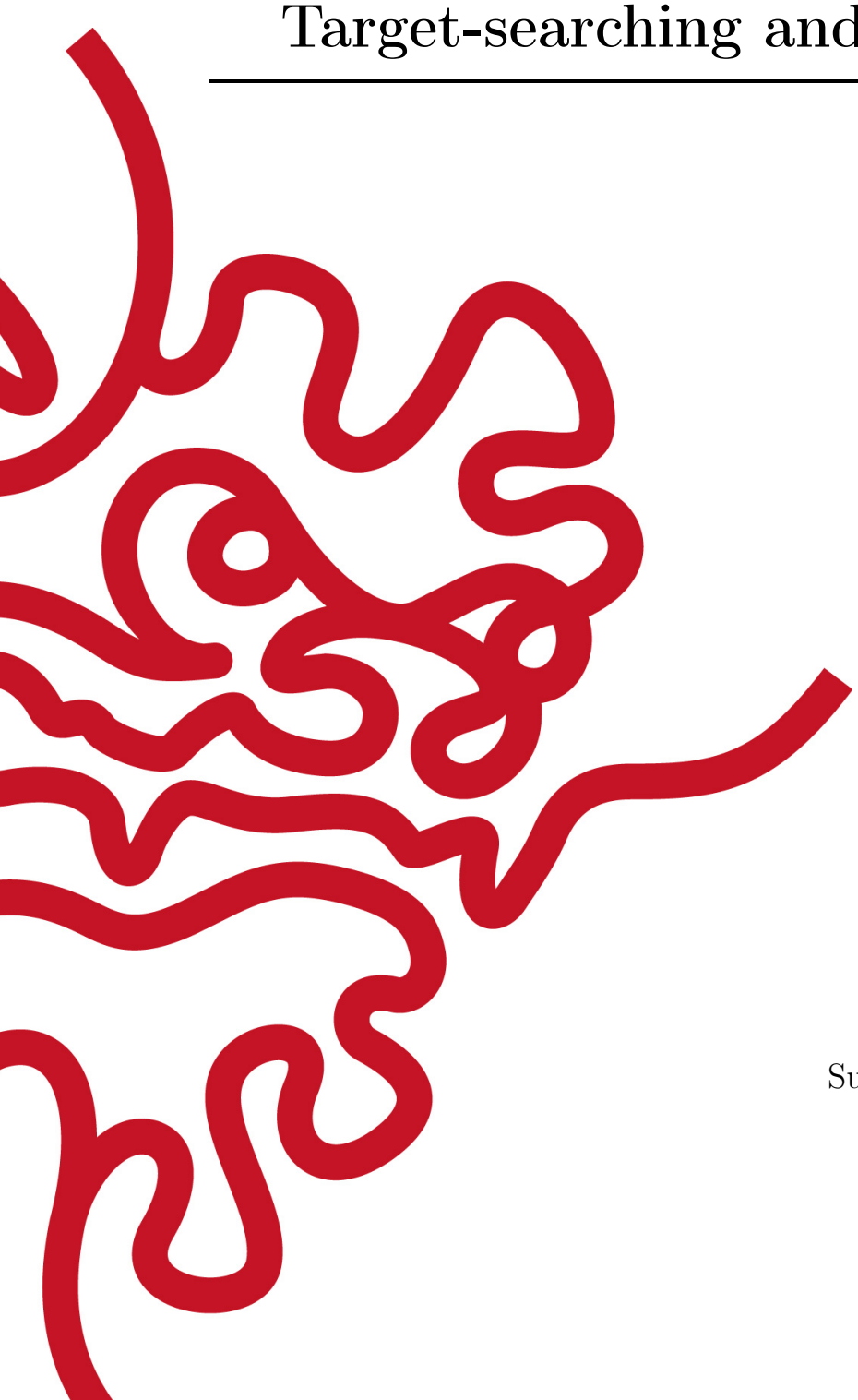
Okinawa Institute of Science and Technology
Graduate University

Thesis submitted for the degree

Doctor of Philosophy

Biophysical Modeling of Cas9 Target-searching and Recognition



by

Qiao Lu

Supervisor: **Simone Pigolotti**

October 2023

Declaration of Original and Sole Authorship

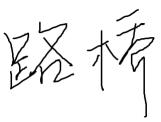
I, Qiao Lu, declare that this thesis entitled *Biophysical Modeling of Cas9 Target-searching and Recognition* and the data presented in it are original and my own work.

I confirm that:

- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly acknowledged. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- None of this work has been previously published elsewhere, with the exception of the following:

  Lu, Q., Bhat, D., Stepanenko, D., Pigolotti, S. *Search and localization dynamics of the CRISPR-Cas9 system*, Physical Review Letters, 2021. In this letter, Bhat, D. computed the data in Fig. 4 (a) (b) (Fig 3.4 (a) (b) in this thesis); Stepanenko, D. wrote the preliminary code in python. The rest (most) of the work were done by me under the instruction of my supervisor.

Date: October 2023
Signature:



Abstract

In this thesis, I model the 1D diffusion and unbinding of Cas9 on/from DNA. Cas9 plays a key role in the CRISPR/Cas (CRISPR-associated protein) system. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) are regions of a prokaryote DNA in which palindromic sequences are interspaced by sequences of foreign origins. These foreign sequences can be transcribed into guide RNAs. Cas9 is an enzyme that combines with guide RNA and afterwards can recognize and cleave DNA strands that are complementary to the guide RNA. However, it is not clear how does Cas9 identify its target. The protospacer adjacent motif (PAM) is a "NGG" segment on DNA. The recognition of PAM is the initial stage of the target searching mechanism. Experiments suggest that Cas9 uses 3D diffusion combined with 1D diffusion along the DNA, a mechanism termed facilitated diffusion. My model explains the distribution of binding events observed in experiments and predicts biophysically relevant parameters. I then analyze the behavior of Cas9 on a generic DNA with disordered assortment of PAMs by using an analogy with Anderson localization in condensed matter physics. I then propose a model of the off-target behavior (specificity) of Cas9. From the measured rates, I determine the energy landscapes of on-target and off-target DNA sequences, and the thermodynamic parameters in double strand DNA and DNA-RNA hybrids. Finally, from a perspective of two-mode target recognition strategy, I investigate the effect of PAM and its binding energy on the efficiency of Cas9 in the facilitated diffusion process.

Acknowledgment

The guidance and help from my supervisor is abundant in both the research and the writing of this thesis. Deepak Bhat computed the data in Fig 3.4 (a) (b); Darya Stepanenko wrote the preliminary code in python. Besides, we (me and my suervisor) are grateful to Chirlmin Joo and Viktorija Globyte for sharing experimental data. We thank Zev Bryant for kind discussion.

Abbreviations

bp	base pair(s)
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
gRNA	guide RNA
LHS	left hand side
MFPT	mean first passage time
ODE	ordinary differential equation
PAM	protospacer adjacent motif
RDR	reversibility-determining region
RHS	right hand side
TF	transcription factor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is a biophysical study of the target-searching and recognition of the Cas9 protein on DNA. This project extends the theory of facilitated diffusion of proteins such as TF (transcription factor) and deepens our understanding of Cas9. The thesis is organized as follows. In chapter 1, I introduce the CRISPR/Cas system and specifically Cas9. I then qualitatively introduce the concept of facilitated diffusion and its possible role in Cas9 searching. Chapter 2 reviews the theory of Anderson localization and facilitated diffusion in a quantitative way. In chapter 3, I introduce my model of Cas9 one dimensional searching. The model explains the distribution of binding events observed in experiments, and predicts biophysically relevant parameters. I then analyze the behaviour of Cas9 on a generic DNA by using an analogy with Anderson localization. Chapter 4 investigates the efficiency of Cas9 from a perspective of a two-mode target recognition strategy. In chapter 5, I propose a model of the off-target behaviour (specificity) of Cas9. Chapter 6 presents my conclusions.

1.1 Introduction to Cas9

1.1.1 The CRISPR-Cas system

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas (CRISPR-associated protein) are the immune systems of prokaryote cells. They were originally found in Escherichia coli [Ishino et al., 1987, Mojica et al., 2000]. A sketch of the CRISPR-Cas system from [Bonomo and Deem, 2018] is shown in fig 1.1. CRISPR are palindromic sequences repeated for several times in the DNA [Jansen et al., 2002]. They are found in approximately 50% of sequenced bacterial genomes and about 90% of sequenced archaea [Hille et al., 2018]. Neighbouring palindromic sequences are interspaced by sequences of foreign origins, usually derived from DNA fragments of bacteriophages that had previously infected the prokaryote [Barrangou et al., 2007]. In contrast to the palindromic sequence, these fragments (spacers) are different from each other. Foreign sequences in CRISPR can be transcribed into RNA. Those short RNA molecules, which are named guide RNA [Brouns et al., 2008], are then combined with Cas proteins [Van Der Oost et al., 2014]. The complex they form is called a ribonucleoprotein (RNP).

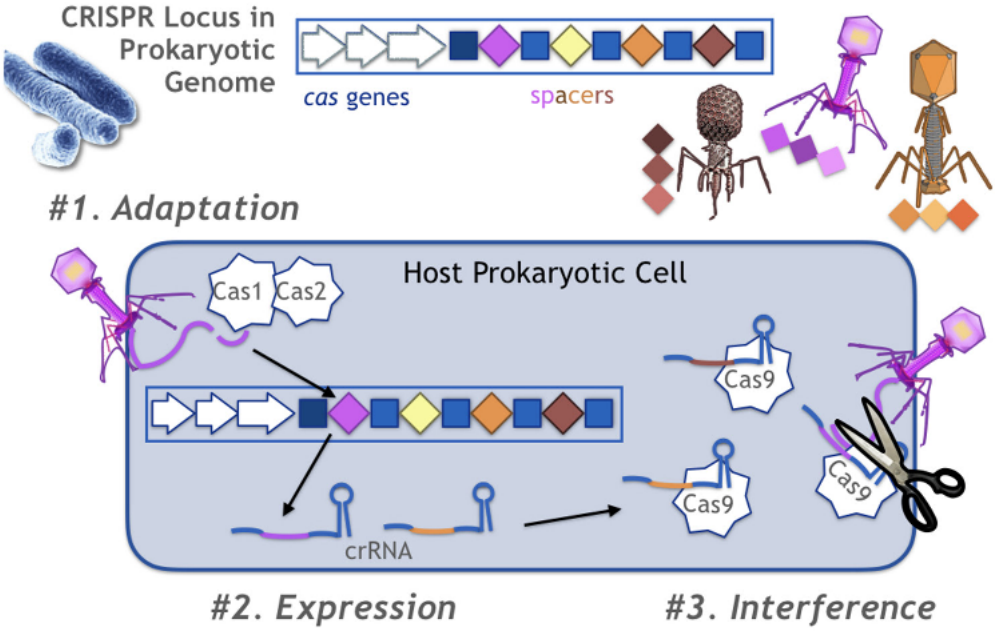


Figure 1.1: The functioning of CRISPR-Cas immune system. Different spacers are shown by different colours, and the blue squares between them are the repeated palindromic sequences. gRNA is composed of crRNA (CRISPR RNA) and tracrRNA (trans-activating crRNA). The latter plays a role in the maturation of the former. These details are omitted in the sketch as well as in the text, and are irrelevant to this thesis. From [Bonomo and Deem, 2018]

Cas is a DNA endonuclease enzyme. RNPs are able to recognize and cleave DNA strands that are complementary to the guide RNA, and therefore damage DNA molecules that are identical to the foreign sequences stored in CRISPR. In this way, CRISPR and Cas act as the immune system in prokaryotes, to identify and destroy invading DNA segments, such as those of viruses [Westra et al., 2012]. Once a bacteria colony is infected by a virus that was not been encountered before, the surviving bacteria would incorporate a DNA fragment from the new virus into CRISPR, and be able to defend themselves therefrom [Fineran and Charpentier, 2012]. Since CRISPR is inherited by daughter cells when bacteria divide, the immune ability for a particular virus is inherited as well. The whole immune process can be divided into 3 stages: adaptation of foreign DNA, expression into gRNA, and interference [Bonomo and Deem, 2018].

1.1.2 A general survey of Cas9

There are 3 major types of Cas proteins, based on their genetic content and structural differences [Makarova et al., 2017a, Makarova et al., 2017b]. The Cas9 found in Streptococcus pyogenes is a prototype in type II. Cas9 assembles a guide RNA (gRNA), and forms a RNP complex which recognizes and destroys the invading genetic segment [Gasiunas et al., 2012]. Its target is a 23 base pair double strand DNA sequence composed of two parts. The first part is a "NGG" sequence called PAM (protospacer adjacent motif), in which the "N" can be any base [Anders et al., 2014]. The second part is a 20 base pair sequence, complementary to the gRNA, that is found upstream of the

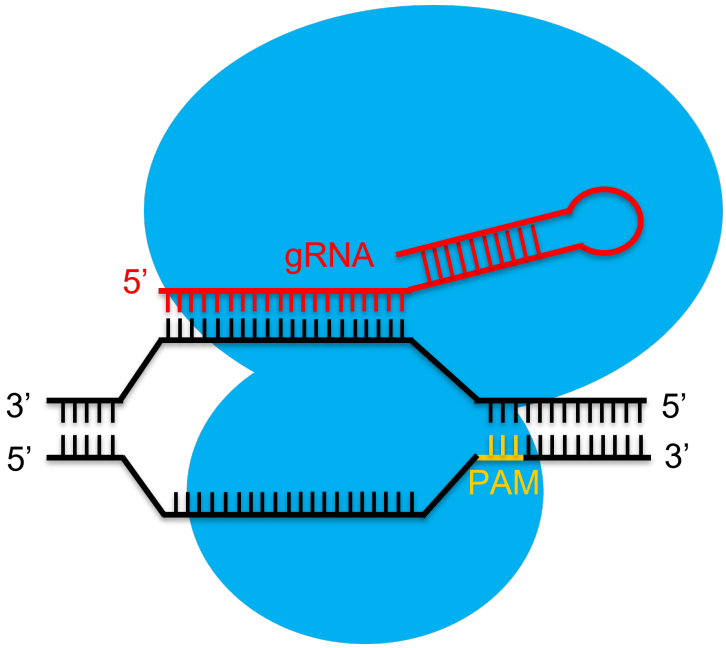


Figure 1.2: Cas9 and its target, with the PAM marked in yellow. The gRNA (in red) is forming a heteroduplex with the target sequence.

PAM [Mojica et al., 2009]. Studies on the target searching mechanism of Cas9 by single molecule method and in bulk show a general picture of on-target binding, in which Cas9 recognizes the PAM first, then DNA melts, followed by formation of the heteroduplex, and finally cleavage [Szczelkun et al., 2014, Martens et al., 2019, Globyte et al., 2019] [Boyle et al., 2017, Sternberg et al., 2014]. A sketch of Cas9 recognizing its target is shown in fig 1.2.

Cas9 can transiently bind to a PAM even in the absence of a neighboring target [Globyte et al., 2019, Jones et al., 2017]. In this case, the dsDNA does not melt, and the binding to the PAM and its neighbouring base pairs lasts about 3 seconds [Globyte et al., 2019]. If the PAM is next to a perfectly matched target, a control experiment in [Sternberg et al., 2014] shows that the nucleating of the RNA-DNA heteroduplex starts from the position next to the PAM, and proceeds sequentially towards the distant end of the target. The control experiment excludes the unsequential possibility that the nucleation of the RNA–DNA heteroduplex starts at any other position of the target farther from PAM. Although sequential, the unwinding process does not proceed at constant speed and can be divided into different stages. [Ivanov et al., 2020] used a single-molecule technique called rotor bead tracking (RBT) to investigate the dynamics of Cas9 R-loop formation and collapse. By recording the DNA unwinding angle and the corresponding number of base pairs, they found that there is a intermediate state between the closed (i.e. unwound) state and the open R-loop state. This intermediate state corresponds to 9 to 10 base pair unwinding. The transitions between this state to/from both the closed state and the open state are instantaneous, compared with the time the DNA spend within these three states. [Gong et al., 2018] also showed that the R-loop formation is a two step process, corresponding to the adjustments of two domains of Cas9.

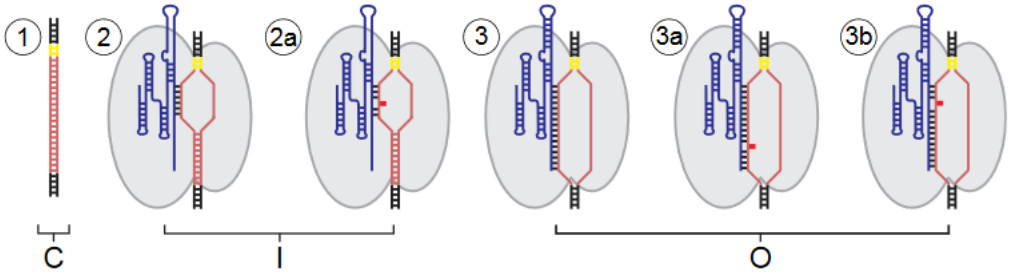


Figure 1.3: The three states observed in [Ivanov et al., 2020]. C, I, O represents closed, intermediate, and R-loop (open) state, respectively. 1, 2, 3 are on-target cases, 2a, 3a and 3b are different off-target cases. The number of base pairs is only schematic. From [Ivanov et al., 2020].

The transition time from PAM binding to formation of the R-loop is hard to measure due to experimental time resolution. But an estimate is within approximately 0.1 second [Ivanov et al., 2020], as also supported by previous studies [Sternberg et al., 2014, Singh et al., 2016, Jones et al., 2017, Singh et al., 2018, Gong et al., 2018].

1.1.3 Specificity of Cas9: observed properties and modelling works

The off-target case, in which the PAM is next to a target but with a few base pairs' mismatches, is key to understand specificity of Cas9. Specificity is important because, besides on-target (perfectly matched) sequence, Cas9 can also bind and cleave some off-target sequences [Boyle et al., 2017, Bonomo and Deem, 2018]. Experiments focusing on the off-target behaviour often use dCas9, which is a variation of Cas9 without the amino acid residue for cleavage.

Important features observed in experiments are listed in the following.

[Ivanov et al., 2020] studied the off-target behaviour for a small set of mismatched sequences. They found that:

- For most of the mismatched sequences, they observed a single intermediate state. Only in a minority of cases, the position of this intermediate state is different, or there are more intermediate states (2 or 3). In on-target and different off-target cases, the transition rates from the closed state/R-loop state to the intermediate state and vice versa are all different. The three states are represented in fig 1.3.

We denote the PAM sequence base pairs by -3, -2 and -1, and "NGG" is called the canonical PAM, with any modified version been called noncanonical PAM. To be consistent with the notation for the PAM, the 20 base pair sequence complementary to the gRNA is marked from 1 to 20, from the base pair next to the PAM to the most distant pair. A figure showing the numbering and different regions defined in this subsection is shown fig 1.4.

[Boyle et al., 2017] measured the apparent association rates, the equilibrated occupancy after incubation, and finally the apparent dissociation rates for all possible 1 or 2 base pair substitutions of the 23 base pair target. In such a way, a library of the

Figure 1.4: The numbering convention of the target. Yellow: PAM, green: seed region, blue: RDR (reversibility-determining region).

dCas9 quantitative behaviour was built for every modified (up to 2 base pair) targets. Important features include:

- **Rates constant in time.** The association/disassociation rates they measured are nearly constant for a long time interval (more than 1500 seconds) for all sequences except for those with a negligible binding rate.
- **Cannonical & noncanonical PAM.** The apparent association rates of targets with noncanonical PAM is low, as observed by previous studies. However, the equilibrated occupancy of the NGA and NAG PAMs are larger than that of other noncanonical PAMs.
- **The seed region.** Investigation of single mismatches in that sequence (but with a canonical PAM) reveals a "seed region" from 1 to 8, where substitutions significantly reduce the apparent association rate. For more distant single mismatches, the decrease is no more than twofold. Similar to the PAM, mismatches at the same position but with different bases lead to different rates. Although the seed region strongly affects the apparent on-rate, mismatches within it do not appreciably decrease the equilibrium occupancy.
- **Nonlinearity in double-mismatches.** By "Nonlinearity" we mean that double mismatches have larger energy barriers for association than the sum of the energy barriers of the corresponding two single mismatches. This effect is more pronounced for sequence beyond the seed region. In particular, if one or both mismatches are in the 8 to 11 region, this nonlinearity is very strong and the association rate is quite low. In general, unlike single mismatches, double mismatches markedly reduce the long-time occupancy.
- **The reversibility-determining region** The dissociation rate is large when there are one or more mismatches in the so-called "reversibility-determining region" (RDR), that is from position 8 to 17. If there are no mismatches in this region, the off rate is quite low. In contrast, the seed region plays a vital role in association/equilibrium case but is of no account here. The time scale was proved to be important for this. When the association time before unbinding was 45 mins rather than more than 10 hours, the PAM and seed region was found to be as important as the RDR region.

The mechanism of Cas9 recognition may be inferred from these observations. A model of strand invasion, in which a single-stranded nucleoprotein filament moves into the similar or identical recipient DNA duplex, should be able to explain the different

role of different regions among the sequence, their synergetic behaviour, different time scales, and other properties.

Each existing model of the specificity of Cas9 has its unique advantages. For example, [Klein et al., 2018] builds a relatively simple model that performs well in general. However, most models are not be able to describe all important features listed above [Josephs et al., 2015, Klein et al., 2018, Khakimzhan et al., 2020, Marklund et al., 2022]. Others are very successful in explaining all of the key features, but have other disadvantages. For example, [Feng et al., 2021] contains ad hoc fitting parameters (a uniform "compensation energy" in every unwinding steps). Another example is the very recent [Eslami-Mossallam et al., 2022] that did not consider that the association/disassociation rates are constant in time, as mentioned earlier in this subsection. Its assumptions are also oversimplifying, for example, for a certain mismatch position, they do not distinguish mismatch base pair types.

1.2 Facilitated Diffusion and Cas9

1.2.1 Introduction to facilitated diffusion

Many proteins such as transcription factors (TF) act by binding to particular sites on the DNA, in such a way to regulate the corresponding gene expression. The facilitated diffusion discussed here is in this context of protein finding their target site on DNA. It was first reported that lacI repressor could find its target site approximately two orders of magnitude faster than predicted by 3D diffusion only (i.e. diffusion in cytoplasm and random collision) [Riggs et al., 1970]. Then, a series of seminal papers [Berg et al., 1981, Winter and Von Hippel, 1981, Winter et al., 1981] proposed that the TF may combine 1D diffusion (sliding) along the DNA with 3D diffusion to locate its target. The sensitivity of the result to salt concentration [Riggs et al., 1970] was interpreted as evidence that the DNA electrostatically attracts the protein, which provided a physical mechanism for sliding. By using realistic parameters from experiments, [Dahirel et al., 2009] theoretically demonstrated that, for sequence specific DNA-binding proteins such as transcription factors and restriction enzymes, sliding by electrostatical force is possible. Besides sliding, other mechanisms such as hopping, i.e, the possibility for proteins to briefly detach from DNA and then reattach at short distance, were also proposed in [Berg et al., 1981].

In order to fulfill its biological function, a TF needs to bind its target site tightly, but from the energetic aspect this requires a large binding energy difference between different sequences. This conflict is termed the paradox of speed and stability in facilitated diffusion. For 1D diffusion, theory shows that a rough energy landscape results in a smaller effective diffusion constant [Zwanzig, 1988], hence reducing the efficiency of sliding. Many possible solutions and mechanisms have been proposed to resolve this paradox [Slutsky and Mirny, 2004, Mirny et al., 2009, Sheinman et al., 2012, Cencini and Pigolotti, 2018].

1.2.2 Cas9 searches its target by facilitated diffusion?

Experimental evidences on whether Cas9 performs facilitated diffusion are inconsistent with each other [Sternberg et al., 2014, Singh et al., 2016, Globyte et al., 2019]. One way of probing the search dynamics of Cas9 is to experimentally measure the distribution of Cas9 molecules bound along the DNA. For example by single-molecule study using DNA curtains [Sternberg et al., 2014]. This work did not observe sliding, but found that Cas9 is localized in regions that extend for hundreds of base pairs length around targets. Another study [Singh et al., 2016] did not find evidence of sliding either.

Single-molecule FRET (Forster resonance energy transfer) technique was used to study how Cas9 interacts with the PAM in [Globyte et al., 2019]. FRET enables monitoring of the real-time position of the protein relative to a specific position on the DNA contour. In the absence of PAM, binding of Cas9 on DNA can only last for a very short time interval before detachment. The number of Cas9-DNA binding events decays exponentially with the binding duration, implying a constant dissociation rate. In the presence of one or more PAM(s), one observes appreciably longer binding events. These long events present another approximately exponential behaviour, with a much longer characteristic time scale, in which the Cas9 is found to be searching the DNA sequence near the PAM, or translocating between PAMs in the multi-PAM case. This phenomenon is termed "double exponential decay" in [Globyte et al., 2019]. The main experimental results are summarized in Fig 1.5 [Globyte et al., 2019]. Along with the observation that Cas9 can move between PAMs in multi-PAM experiments, the results in [Globyte et al., 2019] reveal that Cas9 not only uses 3D diffusion as suggested before [Sternberg et al., 2014], but also 1D diffusion along the DNA. Moreover, the sliding length suggested by [Globyte et al., 2019] is much shorter than the hundreds of base pairs suggested by [Sternberg et al., 2014].

Although conflicting upon whether Cas9 can sliding, the double exponential decay is also observed in [Sternberg et al., 2014] with lifetime of about 3.3s and 58s, and observed in [Singh et al., 2016], too. In relative systems, such as in type I CRISPR-Cas systems, the pause time of Tfu (Thermobifida fusca) complex Cascade on DNA shows a double exponential behaviour as well, with the short and long half life being 1-3s and 50s, respectively [Dillard et al., 2018].

[Hammar et al., 2012] discusses the binding of transcription factors (TF) on lac operators, in which there is a similar phenomenon with results observed for Cas9. There is interference between two lac operators on DNA due to facilitated diffusion. Here "interference" means that the distance between the two operators can change the detachment rates and other observables. This effect can be derived quantitatively [Hammar et al., 2012]. Different distances between neighbouring PAMs also leads to different behaviour of Cas9 [Globyte et al., 2019]. This is another evidence that Cas9 adopts facilitated diffusion.

Similar to the paradox of speed and stability in TF, it is possible that there is also a trade-off in the context of Cas9, i.e. spending too much time to investigate the vicinity of each PAM would slow the searching task, but skimming too quick over PAMs might risk missing the target sequence. This trade-off might exist, regardless of whether Cas9 searches by facilitated diffusion or not.

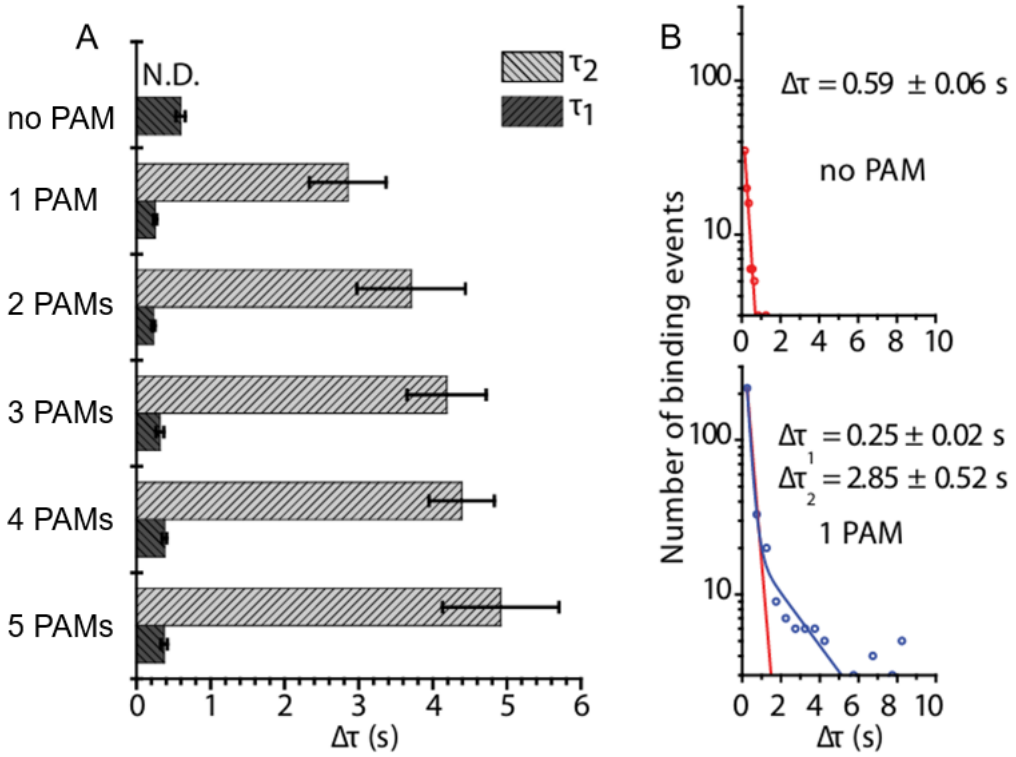


Figure 1.5: From [Globyte et al., 2019]. A: Cas9 dwelltimes (average durations of DNA binding events) with standard errors. Binding events are divided into short and long events (no 20 base pair here). Only short (black) bindings are present when there is no PAM but both short and long (gray) exist when there is PAM, and the length of the long binding event increases with the number of PAMs. B: histogram of binding events by their time interval, both in logarithm plot. The upper panel of no PAM shows a fast exponential decay, implying a constant large detachment rate. The bottom panel of 1 PAM displays a slower exponential decay (the part of the blue curve that diverges from the red), apart from the red fast exponential decay which is still present at short times. This slower decay implies another constant but smaller detachment rate, and hence a double exponential behaviour (the entire blue curve).

Chapter 2

Theoretical preliminaries

In this chapter, I review two existing theories/results that are crucial for my work. The first section is about Anderson localization, that will be used in chapter 3. The second section is about the average searching time in facilitated diffusion and will be used in chapter 4.

2.1 Anderson localization

The idea of Anderson localization is related to the absence of wave diffusion in a disordered medium [Anderson, 1958]. Anderson considered a electron wave function in a lattice in which the potential energy of each site is random. He proved that the electron eigenstates are no longer Bloch functions, and some eigenstates are localized, in the sense that the magnitude of the wave function decays exponentially in space. Localization can also occur in classical systems, for example in the 1D oscillator chain shown in fig 2.1.

The Anderson model considers a 1D lattice, in which the amplitude of the wave function at site n is ψ_n. The wave function satisfies the discretised stationary Schrödinger equation:

$$-\psi_{n+1} - \psi_{n-1} + 2\psi_n + V_n\psi_n = E\psi_n \tag{2.1}$$

in which the first three terms on LHS come from the discretised Laplace operator acting on ψ_n, V_n is the random potential on site n and E is energy. Constants such as the mass and $\hbar$ are rescaled to 1. Then, for a localized eigenstate with its amplitude peaked at n^*, the envelope of its magnitude decays exponentially. This means that [Crisanti et al., 2012]

$$|\psi_n| \leq |\psi_{n^*}|e^{-c|n-n^*|}, \tag{2.2}$$

where c is a constant. The localization length γ is defined as

$$\gamma^{-1} = -\lim_{|n|\to\infty} \frac{1}{|n|}\langle \ln|\psi_n|\rangle, \tag{2.3}$$

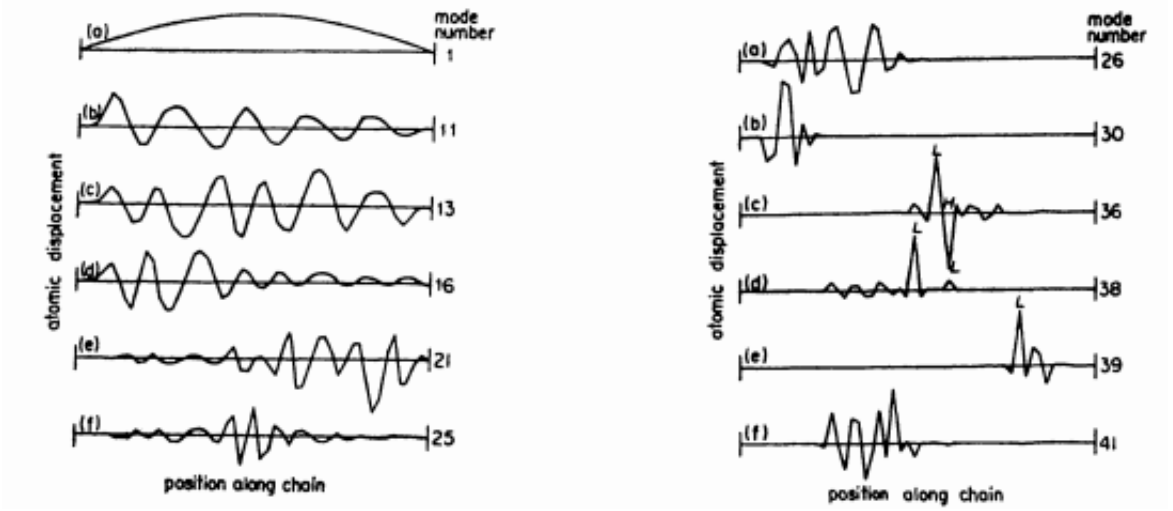


Figure 2.1: Normal modes of a 1D oscillator system, in which 25 light atoms and 25 heavy atoms are connected by identical springs. The left panel shows more extended normal modes with lower frequencies. The right panel shows several strongly localized normal modes with intermediate and high frequencies. From [Ishii, 1973], originally from [Dean and Bacon, 1963].

where the average is taken over the disorder. The position of n^* and the value of $|\psi_{n^*}|$ do not matter because of the $|n| \to \infty$ limit. However, for an arbitrary E and some (left) boundary conditions ψ_0 and ψ_1, one cannot calculate γ by first iterating Eq. (2.1) to obtain ψ_n, then using this definition. To show why this does not work, and also to find a viable alternative, we have to introduce the tool of transfer matrix.

We introduce the vector $\boldsymbol{\psi}_n = (\psi_n, \psi_{n-1})$ and the transfer matrix

$$\hat{T}_n = \begin{pmatrix} 2 + E + V_n & -1 \\ 1 & 0 \end{pmatrix}. \tag{2.4}$$

With these definitions, we rewrite Eq. (2.1) as

$$\boldsymbol{\psi}_{n+1} = \hat{T}_n \boldsymbol{\psi}_n \tag{2.5}$$

and therefore

$$\boldsymbol{\psi}_N = \prod_{n=1}^{N-1} \hat{T}_n \boldsymbol{\psi}_1. \tag{2.6}$$

V_n are mutually independent, identically distributed random quantities, so $\hat{T}_n$ are independent and identically distributed (i.i.d.), symplectic random matrices. Then, the Furstenberg theorem [Furstenberg, 1963, Matsuda and Ishii, 1970, Furstenberg, 1971] [Ishii, 1973] shows that for any nonzero $\boldsymbol{\psi}_1$, the modulus of the vector of site N, $|\boldsymbol{\psi}_N|$ satisfies

$$\lim_{N\to\infty} \frac{1}{N} \langle \ln |\boldsymbol{\psi}_N| \rangle = \Lambda_1 > 0, \tag{2.7}$$

with probability 1, as long as the transfer matrices satisfy some mild conditions: essentially, for common problems in physics, the only key requirement is that in the ensemble of $\hat{T}_n$, there are at least two elements with no common eigenvectors [Ishii, 1973]. Here, Λ_1 is the maximum Lyapunov exponent.

This result implies that, for almost any (left) boundary condition $\boldsymbol{\psi}_1$, ψ_N does not decrease, but increase exponentially with N with a rate Λ_1 (specifying $\boldsymbol{\psi}_1$ is the same as specifying one number only, since only the ratio ψ_1/ψ_0 matters). This is because, for a given ψ_1/ψ_0, a right boundary condition and a given realization of the random sequence V_n, an arbitrary E is not a eigenvalue. For such a given system, only when E is equal to an eigenvalue E_n, the exponential growth of ψ_n starting from the left boundary can match the other exponential growth (with decreasing n) of ψ_n from the right boundary, at some peak position n^* in the middle.

In such a eigenstate, the localization length γ is equal to the inverse of Λ_1 computed by transfer matrices and Eq. (2.7) with $E = E_n$. This statement is called the Borland conjecture [Borland, 1963]. In situations where this conjecture holds, one can numerically calculate γ using transfer matrices:

$$\gamma^{-1} = \Lambda_1(E) = \lim_{N\to\infty} \frac{1}{N} \ln |\boldsymbol{\psi}_N(E)| = \lim_{N\to\infty} \frac{1}{N} \ln \left| \mathrm{Tr} \prod_{n=1}^{N} \hat{T}_n \right|. \tag{2.8}$$

The value of $\Lambda_1(E)$ does not depend on the realization of the disorder and $\boldsymbol{\psi}_1$ because of the Furstenberg theorem.

One way to solve the spectrum of a 1D disordered system is as follows [Herbert and Jones, 1971, Thouless, 1972]. For a fixed ψ_1/ψ_0, ψ_N is a polynomial of degree $N-1$ in E. Therefore

$$\psi_N = A \prod_{n=1}^{N} (E_n - E), \tag{2.9}$$

where E_n are the zeros of ψ_N and A is a normalization factor that does not grow with N. This can be seen easily by calculating the coefficient of E^{N-1}. Then, E_n are the eigenvalues of a chain satisfying the boundary conditions of the fixed value of ψ_1/ψ_0 and $\psi_N = 0$. Each factor on the RHS of Eq. (2.9) can be positive or negative depending on whether E is smaller or larger than E_n, so

$$(E_n - E) = |E_n - E| \mathrm{e}^{i\pi\theta(E-E_n)}, \tag{2.10}$$

where θ is the Heaviside step function. Substitude this expression into Eq. (2.9) we have

$$\frac{1}{N} \ln \psi_N(E) = \frac{1}{N} \sum_{n=1}^{N-1} \ln |E_n - E| + \frac{i\pi}{N} \sum_{n=1}^{N-1} \theta(E - E_n) + \frac{1}{N} \ln A. \tag{2.11}$$

Eq. (2.11) is known as the Herbert-Jones-Thouless formula [Herbert and Jones, 1971]. Another proof by using Green's function is in [Thouless, 1972]. Numerically, one can calculate $\psi_N(E)$ by transfer matrices and then vary E. Wherever $\psi_N(E)$ changes sign, this implies that E has reached an eigenvalue E_n. When N is large, the spectrum is

not sensitive to either the given boundary conditions of ψ_1/ψ_0 and $\psi_N = 0$, or the realization of the sequence $\hat{T}_n$. In the limit $N \to \infty$, it only depends on the probability distribution of $\hat{T}_n$.

2.2 The average time to reach a specific target in facilitated diffusion

This section presents a derivation of the average time to reach a specific target in facilitated diffusion, which will be generalized in my work in chapter 4. The main body of this section comes from [Hachmo and Amir, 2022], and will not be cited repeatedly in this section.

In [Hachmo and Amir, 2022], the DNA chain is modelled as a continuous 1D segment of length $2L$, with position denoted by x, with $-L \leq x \leq L$. An unique target is at the origin. During a 1D search round, a protein bound at any $x \neq 0$ can either detach with a constant rate k, or diffuse to $x - \delta x$ or $x + \delta x$ within time δt. The constant diffusion rate is $D = \frac{(\delta x)^2}{2\delta t}$.

A 1D search round ends with a detachment event. After that, the protein undergoes a 3D diffusion round of duration t_{3D}, where t_{3D} is drawn from a certain probability distribution. Then the protein reattaches, with a equal probability to land on any x. In such a way, 1D and 3D diffusion rounds alternate, until the protein reaches the target at $x = 0$.

The goal here is to calculate the average time $\langle T_0 \rangle$ to reach the target starting from a 1D diffusion, with the initial binding position given by a uniform distribution.

2.2.1 Probability to find the target in a 1D round

We call $p(x)$ the probability to find the target in a 1D round, given that the process started at position x. This is a time independent function that only depends on x.

We consider a protein bound at position x at time t. If it eventually succeeds in reaching $x = 0$ without detachment in this round, then during t to $t + \delta t$, it diffuses to either $x - \delta x$ or $x + \delta x$ but not detaching. The probability of not detaching is $(1 - k\delta t)$, and the probabilities of diffusing to the left and to the right are equal. A recursion relation for $p(x)$ can be written as

$$p(x) = \frac{1 - k\delta t}{2} p(x + \delta x) + \frac{1 - k\delta t}{2} p(x - \delta x). \tag{2.12}$$

Subtracting $p(x)$ and dividing by δx^2 on both sides, in the limit $\delta x \to 0$ we obtain

$$\frac{d^2 p(x)}{dx^2} = \frac{k}{D} p(x). \tag{2.13}$$

Since we assumed that the DNA chain is very long, we can solve it in the infinite 1D space $(-\infty, +\infty)$. Considering that $p(0) = 1$, the solution is

$$p(x) = \exp\left(-\sqrt{\frac{k}{D}}|x|\right). \tag{2.14}$$

2.2.2 The mean first passage time and the mean failed search time

We denote by $T(x)$ the mean first passage time (MFPT) given that the protein started at x. This is the mean time to reach the target in a 1D diffusion round normalized by the number of all trajectories, i.e. those that failed do not contribute to the numerator of T_n, but they are still counted in the denominator. This means that

$$T(x) = \sum_{i=1}^{\mathcal{N}^s} t_{1D,i}(x)/\mathcal{N}, \tag{2.15}$$

where $\mathcal{N}$ is the total number of trajectories starting from x, $\mathcal{N}^s$ the number of successful trajectories, and $t_{1D,i}(x)$ is the time spent by the i-th trajectory that reached $x = 0$. Similarly, the mean failed search time, given the protein started at x, is denoted by $T^f(x)$. This is the mean time of 1D diffusion contributed by trajectories that failed to reach the target before detachment, normalized by the number of all trajectories,

$$T^f(x) = \sum_{i=1}^{\mathcal{N}^f} t^f_{1D,i}(x)/\mathcal{N}, \tag{2.16}$$

where $\mathcal{N}^f$ is the number of failed trajectories, and $t^f_{1D,i}(x)$ is the time spent by the i-th trajectory before detachment. By definition, $\mathcal{N}^s + \mathcal{N}^f = \mathcal{N}$.

Both $T(x)$ and $T^f(x)$ are time-independent functions, but a recursion relation can be written by considering what happens in a time interval δt. For $T(x)$. Similar to Eq. (2.12), the protein diffuse to both the left and right by δx with equal probabilities $\frac{(1-k\delta t)}{2}$. These events contribute to $T(x)$ as $\frac{(1-k\delta t)}{2}T(x-\delta x)$ and $\frac{(1-k\delta t)}{2}T(x+\delta x)$, respectively. Since a time δt has passed, $T(x)$ includes another term δt but decreased by a factor of $p(x)$, because the trajectories that detached (with probability $1-p(x)$) do not contribute to $T(x)$. Therefore we have

$$T(x) = \delta t p(x) + \frac{(1-k\delta t)}{2}\left(T(x+\delta x) + T(x-\delta x)\right). \tag{2.17}$$

By similar rearrangement as in Eq. (2.13), we have

$$\frac{d^2T(x)}{dx^2} - \frac{k}{D}T(x) + \frac{1}{D}p(x) = 0. \tag{2.18}$$

Given the expression of $p(x)$ in Eq. (2.14), and the boundary conditions $T(0) = 0$, $T(\infty) = 0$, the solution to this ODE is

$$T(x) = \frac{|x|}{2\sqrt{kD}} \exp\left(-\sqrt{\frac{k}{D}}|x|\right). \tag{2.19}$$

The recursion relation for $T^f(x)$ follows the same logic as that for $T(x)$, Eq. (2.17). The only change is to replace $p(x)$ by $1 - p(x)$:

$$T^f(x) = \delta t(1 - p(x)) + (1 - k\delta t)\frac{T^f(x + \delta x) + T^f(x - \delta x)}{2}. \tag{2.20}$$

The ODE then follows:

$$\frac{d^2T^f(x)}{dx^2} - \frac{k}{D}T^f(x) + \frac{1}{D}(1 - p(x)) = 0. \tag{2.21}$$

By substituting again Eq. (2.14), the solution is

$$T^f(x) = -\frac{|x|}{2\sqrt{kD}} \exp\left(-\sqrt{\frac{k}{D}}|x|\right) + \frac{1}{k}\left[1 - \exp\left(-\sqrt{\frac{k}{D}}|x|\right)\right]. \tag{2.22}$$

2.2.3 The expression for $\langle T_0 \rangle$

We denote the starting position of the i-th 1D diffusion as x_i. We introduce the following notations:

- The time spent in the successful i-th 1D search is $t_{1D}(x_i)$.
- The time spent in a failed i-th 1D search is $t^f_{1D}(x_i)$.
- The time spent in the 3D search after the i-th failed 1D search is t^i_{3D}.

These quantities are stochastic.
We denote the inital 1D diffusion by subscript 0. Then the total search time T_0 is:

$$\begin{aligned} T_0 &= p(x_0)t_{1D}(x_0) + (1 - p(x_0))p(x_1) \times \left(t^f_{1D}(x_0) + t_{1D}(x_1) + t^0_{3D}\right) \\ &+(1 - p(x_0))(1 - p(x_0))p(x_1) \times \left(t^f_{1D}(x_0) + t^f_{1D}(x_1) + t^0_{3D} + t^1_{3D}\right) + ... \\ &= \sum_{i=0}^{\infty} p(x_i)\left(t_{1D}(x_i) + \sum_{j=0}^{i-1}\left(t^f_{1D}(x_j) + t^j_{3D}\right)\right) \times \prod_{j=0}^{i-1}(1 - p(x_j)). \end{aligned} \tag{2.23}$$

We now take the mean on both sides. The mean $\langle ... \rangle$ is over the position x as well as over stochasticity. All 1D diffusion rounds start from a uniformly distributed

position. Therefore, for any quantity a in the expression, we have $\langle a(x_i)\rangle = \langle a(x_j)\rangle$, so we can drop the subscipt associated with a specific starting position and write it just as $\langle a(x)\rangle$ (after the $\langle ...\rangle$ operation, it does not depend on x). Note also that different 1D/3D rounds are independent. We then arrive at the following expression:

$$\begin{aligned}\langle T_0\rangle = \sum_{i=0}^{\infty}\Big(\langle p(x)t_{1D}(x)\rangle\left(1-\langle p(x)\rangle\right)^i + i\,\langle p(x)\rangle\left\langle t_{1D}^f(x)\left(1-p(x)\right)\right\rangle\left(1-\langle p(x)\rangle\right)^{i-1} \\ +i\,\langle p(x)\rangle\,\langle t_{3D}\rangle\left(1-\langle p(x)\rangle\right)^i\Big) \\ = \frac{\langle p(x)t_{1D}(x)\rangle}{\langle p(x)\rangle} + \frac{\left\langle (1-p(x))t_{1D}^f(x)\right\rangle}{\langle p(x)\rangle} + \langle t_{3D}\rangle\,\frac{1-\langle p(x)\rangle}{\langle p(x)\rangle}.\end{aligned} \tag{2.24}$$

The quantity $\langle p(x)\rangle = \frac{1}{2L}\int_{-L}^{L} p(x)dx$ can be calculated using Eq. (2.14). $\langle p(x)t_{1D}(x)\rangle$ and $\left\langle (1-p(x))t_{1D}^f(x)\right\rangle$ are by definition $\frac{1}{2L}\int_{-L}^{L} T(x)dx$ and $\frac{1}{2L}\int_{-L}^{L} T^f(x)dx$, respectively. So by using Eqs. (2.19) and (2.22), the final expression is

$$\begin{aligned}\langle T_0\rangle &= \frac{L}{\sqrt{kD}(1-e^{\sqrt{\frac{k}{D}}L})} - \frac{1}{k} + \langle t_{3D}\rangle\left(\sqrt{\frac{k}{D}}\frac{L}{1-e^{\sqrt{\frac{k}{D}}L}} - 1\right) \\ &\approx \frac{L}{\sqrt{kD}} + \langle t_{3D}\rangle\, L\sqrt{\frac{k}{D}}\end{aligned} \tag{2.25}$$

In this expression, we neglected terms that are not proportional to L, assuming that L is very large.

Chapter 3

Interaction of Cas9 with PAM

My own work begins in this chapter. Most of this chapter is taken from my paper *Search and localization dynamics of the CRISPR-Cas9 system* [Lu et al., 2021] apart from this introduction part. There are only some minor changes, in order to fit it into the context of this thesis. Appendices A to D of this thesis are similarly taken from the supplements of [Lu et al., 2021]. In this project, we consider the interaction of Cas9 with PAM, without the 20 bp main target. First, we show that a facilitated diffusion model quantitatively explains the dynamics of Cas9 observed in single-molecule experiments. We then introduce a mapping between facilitated diffusion and Anderson localization. This approach permits us to determine the localization length of Cas9 on typical long DNA strands and explain the discrepancy between the sliding length in [Globyte et al., 2019] and the localization length in [Sternberg et al., 2014] in terms of a hopping mechanism. The mapping presented in this chapter can be used to study the dynamics of other DNA binding proteins, such as transcription factors.

3.1 Equally spaced PAMs

We consider a Cas9 protein that binds on a DNA chain of length N and slides along it before detaching, see Fig. 3.1. Our aim is to quantify the distribution of duration of binding events depending on the arrangement of specific PAM sites along the DNA chain. We introduce the probability $p_n(t)$ that Cas9 is bound at site n at time t, given that it had attached on the DNA at time $t = 0$. Each site represents a nucleotide position $n = 1 \dots N$. We assume attachment to be non specific, so that $p_n(t = 0) = 1/N$.

We distinguish between two types of DNA sites. PAM sites are those at the beginning of a PAM sequence, where Cas9 can bind specifically. We consider every other site as non-specific, including the two other base pairs constituting a PAM, see Fig. 3.1. We call E_n the binding energy of Cas9 at position n. We assume that all non-specific sites have the same binding energy $E_n = 0$. If n is a PAM site, then $E_n = -\beta$, with $\beta > 0$. All energies are expressed in units of $k_\mathrm{B}T$, where k_B is the Boltzmann constant and T the temperature. Our aim is to analyze single binding events and therefore we do not consider rebinding after detachment.

The probabilities $p_n(t)$ evolve according to the master equation

$$\frac{d}{dt}p_n(t) = D_{n,n+1}p_{n+1} + D_{n,n-1}p_{n-1} - (D_{n+1,n} + D_{n-1,n} + k_n)p_n, \tag{3.1}$$

in which $D_{n,m} = De^{E_m}$ and $k_n = ke^{E_n}$, where the diffusion rate D and the unbinding rate k are given parameters. We impose vanishing fluxes at the boundaries, $D_{0,1} = D_{1,0} = D_{N,N+1} = D_{N+1,N} = 0$. This choice of rates satisfies the detailed balance condition $D_{n,m}e^{-E_m} = D_{m,n}e^{-E_n}$.

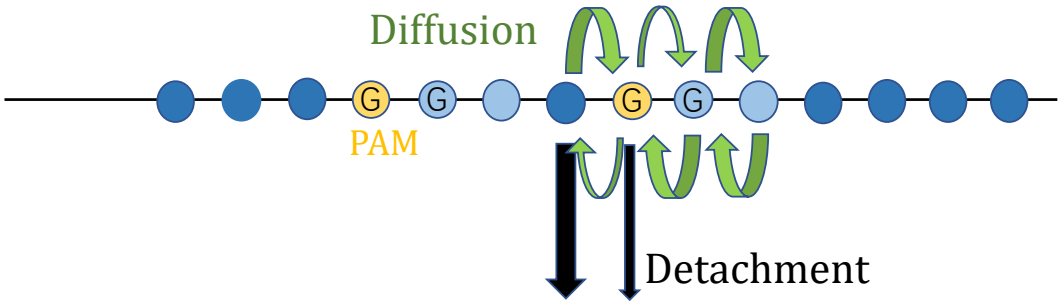


Figure 3.1: Scheme of the model. PAM sites and non-specific sites are shown in yellow and blue, respectively. The second and third bases of PAM sequences are considered as non-specific sites (light blue). Green arrows represent sliding rates and black arrows represent unbinding rates, see Eq. (3.2). Thicker arrows correspond to larger rates.

We express the model in vector notation by defining $\mathbf{p}(t) = (p_1(t), p_2(t) \dots p_N(t))$. We write Eq. (3.1) as $d\mathbf{p}/dt = \hat{A}\mathbf{p}$, where the elements $A_{m,n}$ of the matrix $\hat{A}$ are given by

$$A_{m,n} = \begin{cases} De^{E_n} & \text{if } |n-m| = 1 \\ -(k+2D)e^{E_m} & \text{if } n = m. \end{cases} \tag{3.2}$$

The formal solution to the master equation is $\mathbf{p}(t) = e^{\hat{A}t}\mathbf{p}(0)$, where $\mathbf{p}(0)$ is the uniform initial condition. The eigenvalue equation associated with the master equation is

$$\hat{A}\boldsymbol{\psi} = -\lambda\boldsymbol{\psi}. \tag{3.3}$$

Equation (3.3) is solved by a set of eigenvalues $\lambda = \lambda_1, \lambda_2, \dots \lambda_N$ and associated right eigenvectors $\boldsymbol{\psi} = \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \dots \boldsymbol{\psi}^{(N)}$, assumed to be normalized. The solution of the master equation can be decomposed into eigenvalues

$$\mathbf{p}(t) = \sum_{i=1}^{N} e^{-\lambda_i t} c_i \boldsymbol{\psi}^{(i)}, \tag{3.4}$$

where the coefficients c_i are determined by the initial condition. Because of detachment, one has $\lim_{t\to\infty} p_i(t) = 0$ for all i. This fact and the detailed balance condition imply that all eigenvalues must be real and positive. We sort the eigenvalues so that λ_1 is the smallest one.

The total probability that Cas9 is still bound at a time t is given by $P(t) = \sum_n p_n(t)$. Since we are considering a single binding event, $P(t)$ is a decreasing function of t. We define the instantaneous detachment rate $g(t) = -d/dt\, P(t)$. Single-molecule experiments [Sternberg et al., 2014, Singh et al., 2016, Globyte et al., 2019] observed that

the temporal decay of $g(t)$, and therefore of $P(t)$, is characterized by two distinct exponential slopes at short and long times.

To understand these two regimes, we focus on $P(t)$ and define its instantaneous exponential slope $K(t) = -d/dt \ln P(t)$. We also define the total probability $P_{\mathrm{PAM}}(t) = [\sum_{n\in\mathrm{PAM}} p_n(t)]/P(t)$ of Cas9 being bound to a PAM site at time t, given that it had not detached yet. By summing Eq. (3.1) over n, we find that

$$K(t) = k\,[1 - P_{\mathrm{PAM}}(t)] + ke^{-\beta} P_{\mathrm{PAM}}(t). \tag{3.5}$$

Considering that $0 \leq P_{\mathrm{PAM}}(t) \leq 1$, the slope $K(t)$ is limited by the two unbinding rates:

$$ke^{-\beta} \leq K(t) \leq k. \tag{3.6}$$

The value of $K(t)$ in this range is determined by $P_{\mathrm{PAM}}(t)$. Since the initial distribution is uniform, at short times P_{PAM} is equal to the fraction of PAM sites. Given that this fraction is usually small, Eq. (3.5) implies $K(t) \approx k$ at short times. In the long time limit, Eq. (3.4) leads to conclude that $K(t) = \lambda_1$.

Experiments in [Globyte et al., 2019] measured the distribution of Cas9 binding events on DNA sequences containing from 0 to 5 PAM sites . We jointly fitted the parameters k, β, and D to these six experiments, see Fig. 3.2a. Solutions of the master equation (3.1) with the best-fit parameters reproduce the double exponential behavior and fit well the experimental data, see Fig. 3.2b. The fitted values of the parameters are $k = 1.94 \pm 0.10\, s^{-1}$, $\beta = 3.34 \pm 0.07$, and $D = 52 \pm 9\, \mathrm{bp}^2 s^{-1}$. Experiments on a different variant of Cas9 find differences in binding energy between PAM and near-cognate sites that are comparable with our estimate of β [Farasat and Salis, 2016]. A more detailed model where each non-PAM sequence is characterized by a different binding energy, leads to similar fitted values of the corresponding rates, see appendix. These evidences support robustness of our results.

At increasing number of PAM sites, the second slope in Fig. 3.2a becomes significantly less steep than the first. According to Eq. (3.5), this means that, at long times, Cas9 is much more localized on PAM sites compared with short times. Inspecting the eigenvectors $\boldsymbol{\psi}^{(1)}$ associated with the smallest eigenvalue λ_1 confirms this idea, see Fig. 3.2c.

To gain further insight into the dynamics observed in Fig. 3.2b, we analytically compute λ_1 and its eigenvector for an infinitely long chain with a single PAM site at $n = 0$. For $|n| > 1$, the eigenvector satisfies

$$-\lambda_1 \psi_n^{(1)} = D\left(\psi_{n+1}^{(1)} + \psi_{n-1}^{(1)} - 2\psi_n^{(1)}\right) - k\psi_n^{(1)}. \tag{3.7}$$

We assume a solution of the form $\psi_n \propto e^{-|n|/\ell}$ where $|n| > 0$ and we define ℓ as the sliding length. This solution is inspired by the continuous approximation to Eq. (3.7). Substituting into Eq. (3.7) we obtain

$$k - \lambda_1 = 2D[\cosh(1/\ell) - 1]. \tag{3.8}$$

By expanding the cosh at first order we find $\ell \approx \sqrt{D/(k - \lambda_1)}$. This is expected

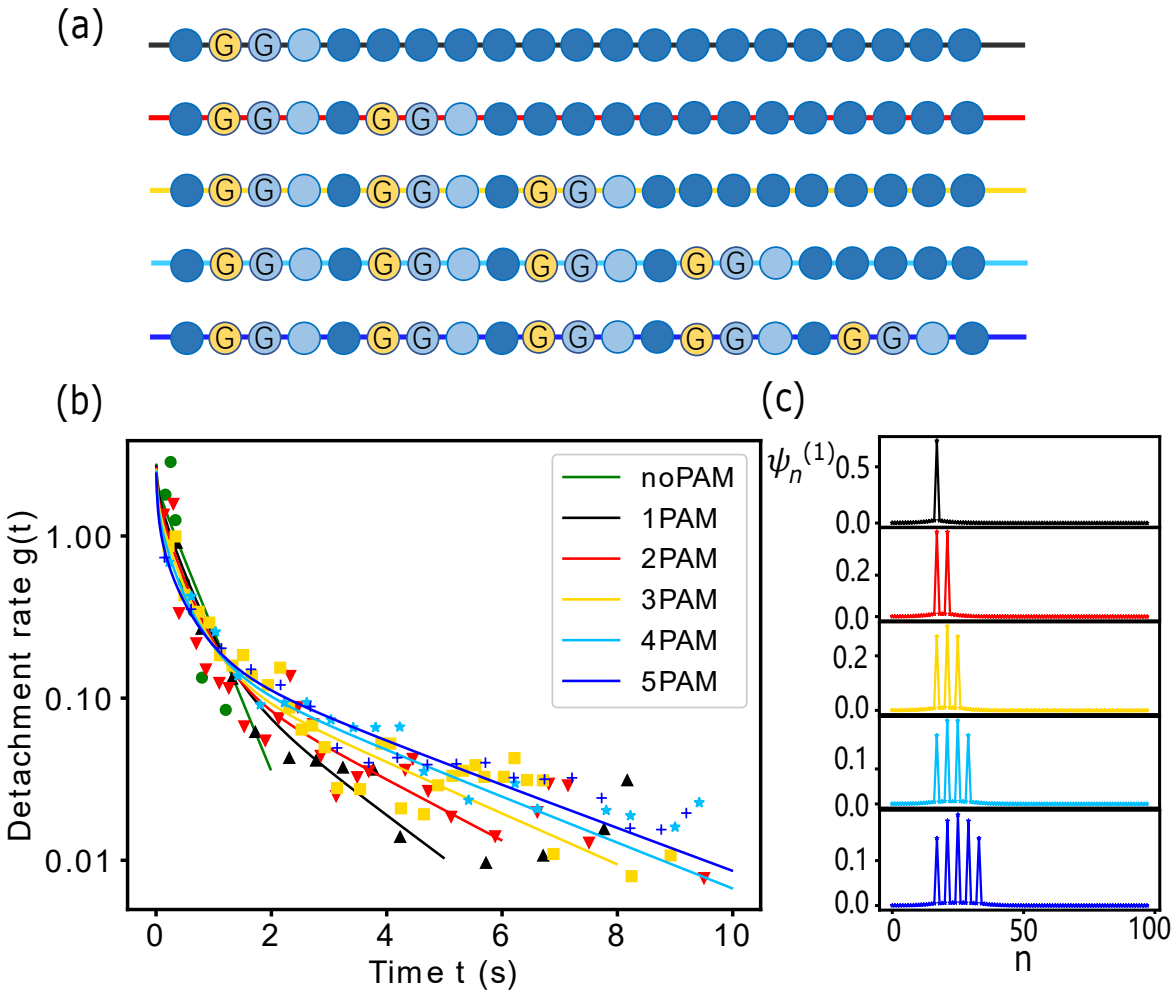


Figure 3.2: (a) Arrangements of PAM sites used in the experiments in [Globyte et al., 2019]. Line colors correspond to the different curves in panel b. The figure shows only the portion of the DNA sequence of length $N = 98$ where the PAM sites are located. (b) Comparison of the prediction of our model (lines) with experiments [Globyte et al., 2019] (points). Model parameters are determined by jointly fitting the experimental data for $j = 0 \ldots 5$ PAM sites using maximum likelihood, see Appendix A. (c) Eigenvectors $\boldsymbol{\psi}^{(1)}$ for $j = 1 \ldots 5$.

because $\psi_n \propto e^{-\sqrt{\frac{k-\lambda_1}{D}}|n|}$ is exactly the solution to the continuous approximation of Eq. (3.7). Note that $\lambda_1 \leq k$ due to Eq. (3.6). The three unknown λ_1, ℓ and ψ_0 (essentially the ratio $\psi_0/\psi_{\pm 1}$) can be determined from Eq. (3.8) and the equivalents of Eq. (3.7) for $n = 0$ and $|n| = 1$. Substituting the fitted parameters of Fig. 3.2, we find $\ell \approx 6.2$ bp. The relative error of λ_1 obtained from this analytical solution to that of the numerical result is about 0.5%. Besides, the $\boldsymbol{\psi}^{(1)}$ from the analytical solution is indistinguishable from the numerical one except at the two ends of the DNA chain. In fact in the continuous limit and for a infinitely long chain the analytical solution will be exact.

Both our model and experiments [Globyte et al., 2019] show that the lifetime of long binding events increases at increasing number of PAMs, see Fig. 3.2b. In the model,

this means that λ_1 is a decreasing function of the number of PAMs. This effect can be explained by interference among PAM sites, i.e. the fact that the eigenvector $\boldsymbol{\psi}^{(1)}$ for j PAM sites is not simply a superimposition of j single-PAM eigenvectors, unless the interval between the PAM sites is much larger than ℓ. Only in this limit binding events around each PAM site behave independently, and the long-time exponential slope becomes independent of the number of PAM sites, see Fig. 3.3. At shorter intervals, interference leads to an increase in target occupancy. This implies that, at large t, $P_{\mathrm{PAM}}(t)$, and therefore the typical lifetime of binding events $1/\lambda_1$, are decreasing functions of the interval between the PAM sites, see Eq. (3.5) and Fig. 3.3.

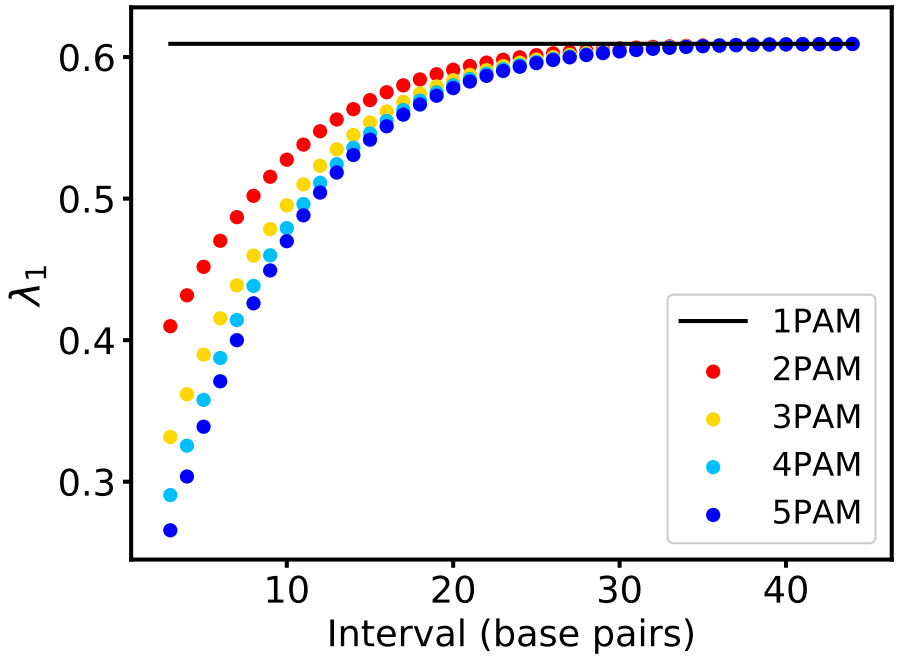


Figure 3.3: Interference between $j = 1 \ldots 5$ equally spaced PAM sites on an infinite DNA chain. Lowest eigenvalue λ_1 as a function of the interval between the PAM sites. Points are obtained by numerically diagonalizing the matrix $\hat{A}$ corresponding to each case, with $N = 220$. The horizontal line marks the value of λ_1 for a single PAM sequence, from the solution of Eq. (3.7).

In summary, we found that the distribution of a Cas9 molecule in a region of DNA containing several PAM sites tends to be localized.

3.2 Generic DNA and disordered assortment of PAMs

We now study the behavior of Cas9 on a very long stretch of DNA including a disordered assortment of PAM sites. The theory of Anderson localization predicts that, in such disordered one-dimensional systems, eigenvectors are exponentially localized:

$$\psi_n \sim e^{-\frac{|n-n^*|}{\gamma(\lambda)}}, \tag{3.9}$$

in the limit $n \to \infty$. Here n^* is the location of the eigenvector peak as in chapter 2 and $\gamma(\lambda)$ is the localization length associated with the eigenvalue λ. Unlike wave functions, here ψ_n is real so there is no need to take the modulus. The localization length γ can be thought as the generalization of the sliding length ℓ: the former is defined for an arbitrary disordered DNA chain and for the envelope of a eigenvector, whereas the latter is defined for a single target. Our hypothesis is that the localization length associated with the smallest eigenvalues of Cas9 dynamics can explain the results of DNA curtains experiments [Sternberg et al., 2014].

We sharpen the analogy between our problem and the Anderson localization by rescaling the components of our eigenvectors by the Boltzmann weight, $f_n = \psi_n \exp(E_n)$. With this transformation, Eq. (3.3) assumes the same form for PAM and non-PAM sites:

$$f_{n+1} + f_{n-1} - \left(2 + \frac{k - \lambda e^{-E_n}}{D}\right) f_n = 0. \tag{3.10}$$

This equation is formally similar to the discrete Schrödinger equation (2.1). It can be solved by the transfer matrix method. We introduce the vector $\mathbf{f}_n = (f_n, f_{n-1})$ and the transfer matrix takes the form

$$\hat{T}_n = \begin{pmatrix} 2 + \frac{k-\lambda e^{-E_n}}{D} & -1 \\ 1 & 0 \end{pmatrix}. \tag{3.11}$$

With these definitions, we rewrite Eq. (3.10) as

$$\mathbf{f}_{n+1} = \hat{T}_n \mathbf{f}_n \tag{3.12}$$

and therefore

$$\mathbf{f}_N = \prod_{n=1}^{N-1} \hat{T}_n \mathbf{f}_1. \tag{3.13}$$

We assume that, in a typical long DNA sequence, each site n has a probability 1/16 to be a PAM site, thereby affecting the value of E_n in the corresponding matrix T_n. As explained in chapter 2, Eq. (3.13) expresses the solution of the eigenvalue equation as a product of random matrices.

The localization length γ can be calculated from this product with Herbert- Jones-Thouless forumula. In this situation Eq. (2.11) and (2.8) reads

$$\frac{1}{N} \ln f_N(\lambda) = \frac{1}{N} \sum_{n=1}^{N-1} \ln |\lambda_n - \lambda| + \frac{i\pi}{N} \sum_{n=1}^{N-1} \theta(\lambda - \lambda_n) + \frac{1}{N} \ln A \tag{3.14}$$

and

$$\Lambda_1(\lambda) = \lim_{N\to\infty} \frac{1}{N} \ln f_N(\lambda) = \lim_{N\to\infty} \frac{1}{N} \ln \left(\mathrm{Tr} \prod_{n=1}^{N} \hat{T}_n \right). \tag{3.15}$$

The transfer matrices for a PAM site and a non-PAM site do not share eigenvector, so the Furstenberg theorem is satisfied, then $\Lambda_1(\lambda)$ is independent of the realization of the disorder and of the choice of $\mathbf{f_1}$.

The validity of the Borland conjecture for our class of systems is supported by numerical and theoretical studies [Ishii, 1973, Matsuda and Ishii, 1970]. Therefore, the inverse of the real part of $\Lambda_1(\lambda)$ can be identified with the localization length $\gamma(\lambda)$. Further, Eq. (3.14) links the imaginary part of Λ with the cumulative density of states. Computing $\Lambda_1(\lambda)$ from the product of transfer matrices, we find that the localization length for the whole spectrum is always shorter than 11 base pairs, see Fig. 3.4b.

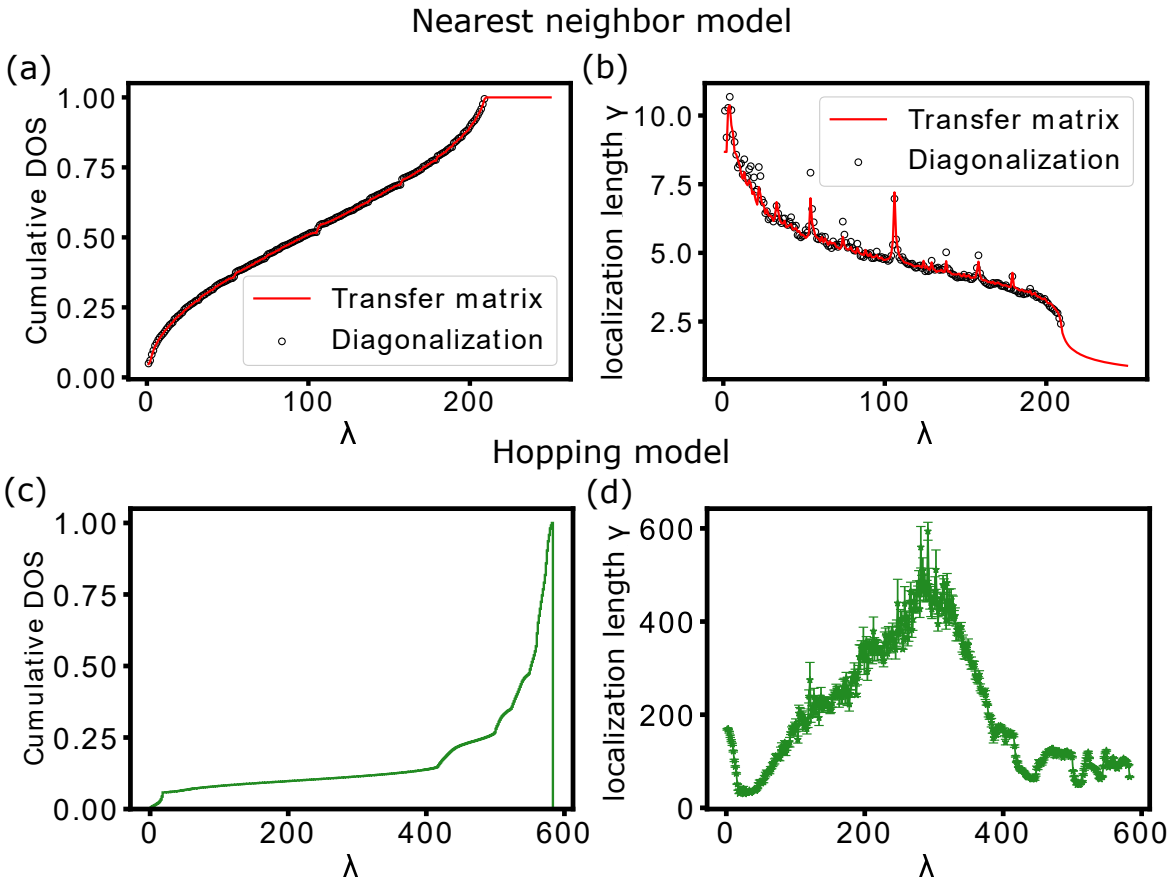


Figure 3.4: (a) Cumulative density of states (DOS) and (b) localization length as function of λ for the nearest neighbour model, Eq. (3.1), computed using Eq. (3.14). Results obtained by the transfer matrix method agree with those obtained by direct diagonalization. The DNA chain length is $N = 10^6$ for the transfer matrix method and $N = 5000$ for the direct diagonalization. (c) Cumulative DOS and (d) localization length for the hopping model expressed by Eq. (3.16), computed using Eq. (3.17). In this case, the DNA chain length is $N = 2000$.

We remark that the disordered arrangement of PAM sites is crucial for this result. In a long DNA chain containing a periodic arrangement of PAM sites, the eigenvectors are extended rather than localized, see Appendix C.

The localization lengths in Fig. 3.4 are much shorter than those observed in DNA curtains experiments [Sternberg et al., 2014]. We assume that this discrepancy can be explained by the following idea. Measuring the distribution of Cas9 in an experiment amounts to performing an "ensemble average" which is potentially affected by search

mechanisms other than sliding (such as hopping). In contrast, FRET experiments focus on individual sliding events, which are unaffected by such mechanisms.

To test this idea, we generalize our model to include hopping. In a hopping event, Cas9 detaches and then reattaches to the DNA at a short distance. This amounts to include in our master equation diffusion among non-nearest neighboring sites:

$$D_{m,n} = De^{E_n} h(|n-m|), \tag{3.16}$$

where $h(n)$ is a positive decreasing function characterizing the probability of hopping events at a given distance n relative to sliding events. We impose $h(1) = 1$, so that nearest-neighbor sliding is consistent with Eq. (3.2). We determine the function $h(n)$ from the solution of a diffusion equation in cylindrical coordinates, see [Lomholt et al., 2009] and Appendix D. Unbinding rates in the hopping model are the same as in Eq. (3.2). For models with next to nearest neighbor interactions, such as our hopping model, the localization length can not be computed using Eqs. (3.14) and (3.15), see [Biddle et al., 2011]. We therefore estimate the localization length by a more direct strategy, although computationally heavier. Assuming that a given eigenvector $\boldsymbol{\psi}^{(i)}$ associated with an eigenvalue λ_i is localized, we obtain from Eq. (3.9) that

$$\gamma(\lambda_i) \sim -\frac{(N-1)}{\ln\left[\psi_1^{(i)}\psi_N^{(i)}\right]}. \tag{3.17}$$

In this case, the localization length associated with the lowest eigenvalues is on the same order of the experimentally measured one (hundreds of base pairs, see Fig. 3.4d).

In conclusion, in this chapter we studied the search dynamics of Cas9 along the DNA. We have shown that the predictions of a facilitated diffusion model with a short sliding length are consistent with the result of single-molecule FRET experiments. By applying the theory of Anderson localization, we have argued that a hopping mechanism can explain how Cas9 is generically distributed along the DNA.

The mapping to Anderson localization introduced in this chapter is a powerful tool that can be applied to any protein performing facilitated diffusion, such as transcription factors.

Chapter 4

The total search time in a motif-guided search mechanism

The work in this chapter will form the basis of a future manuscript. We consider the recognition of the 20 bp target complementary to the gRNA in the context of facilitated diffusion of Cas9. Our focus is to estimate the mean total search time until recognition, and how the energy and density of PAMs affect it. We will see that the trade-off discussed at the end of chapter 1 is indeed present in this context. That is, being too fast when scanning PAMs might risk missing the target sequence, hence longer total search time. But spending more time on the vicinity of each PAM also leads to longer total search time.

4.1 The model with the recognition mode

As in chapter 3 we model the DNA as a discrete lattice. The Cas9 can detach and diffuse with rates

$$\begin{aligned} k_n &= k\exp(E_n) \\ D_{m,n} &= D\exp(E_n), |m-n| = 1 \end{aligned} \tag{4.1}$$

We assume the total length of the DNA is $2L+1$ bp such that $n \in [-L; L]$, with a unique target sequence complementary to the gRNA. At the two boundaries $n = \pm L$, the boundary condition is $D_{\pm(L+1),\pm L} = 0$. The PAM next to the unique target is located at the origin $n = 0$. Since we now consider a generic disordered DNA, for any site $n \neq 0$, the energy E_n is a random variable draw from a probability distribution independent of n. This is our homogeneous assumption for Cas9 in the sliding mode. This assumption does not hold in more complicated cases, e.g. for TFs, there can be a energy funnel around the target [Cencini and Pigolotti, 2018]. A given DNA chain corresponds to a given sequence of E_n, representing the energy landscape of that DNA for Cas9.

We call the state of successfully forming a hybrid of gRNA and the target sequence as the recognition mode (R-mode). The state in which the Cas9 can diffuse and detach is the sliding mode. The Cas9 can only transit to the R-mode from the origin, with a rate $f^T(E_0)$ (T for target). $f^T(E_0)$ is a functions of E_0 only. Once the R-mode is

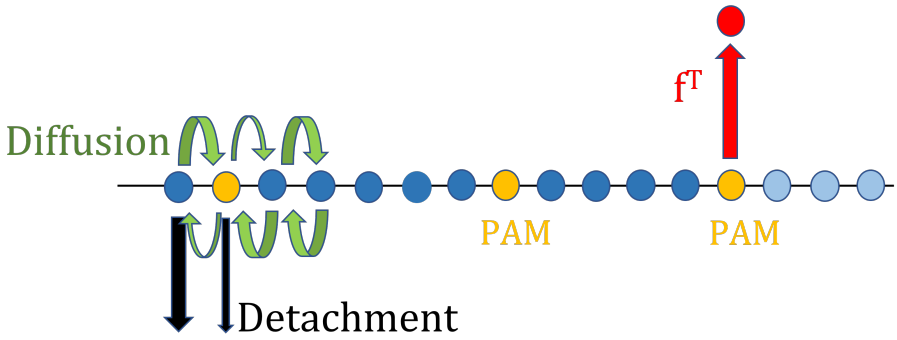


Figure 4.1: A sketch of the model. Circles represent states of the Cas9. Blue and yellow circles represent base pairs on the DNA, and yellow circles represent the starting base pair of a PAM. These are the same as in Fig. 3.1. In contrast, here light blue circles represent the 20 bp target sequence.

reached, the search process is successfully completed. A sketch of the model is shown in Fig. 4.1

Cas9 that detaches before finding the unique target would then do a 3D diffusion round, with duration given by t_{3D}. This is a stochastic quantity draw from a given distribution. After the 3D diffusion round, Cas9 lands on any site with equal probability, and starts its 1D search again.

In this chapter we do not consider hopping. The first reason is that there is no clear cut between a short detaching before reataching and a hopping event. In this type of process, the geometry of DNA can play an important role: in a DNA curtain experiment, DNA chains are straightened, therefore the Cas9 is very likely to reattach to a nearby site relative to its detaching position. In contrast, in a in vivo searching scenario, the DNA chain is rather free so as a first approximation it is reasonable to assume that the rebinding site is randomly distributed. In such a way, the hopping mechanism is at least partly included in the model stated above. Secondly, focusing on diffusion can keep the physics relatively simple, without losing the essential mechanisms.

4.1.1 Model parameters

We estimate $f^T(E_0)$ where E_0 equals the energy of a NGG PAM from experimental results in [Ivanov et al., 2020]. These results show that a R-loop forms approximately 100ms after Cas9 binding, so we take $f^T(E_0) = 10s^{-1}$. To directly measure this rate is hard, because this timescale is beyond experiment time resolution, e.g., 0.1 second for FRET.

$f^T(E_0)$ is the only additional parameter we need for the model. It should satisfy $f^T(E_0) \propto \exp(\Delta G)$, where ΔG is the energy difference of the PAM state and the R-loop state in units of k_BT. But changing E_0 possibly only changes the starting energy baseline of the melting process, and does not necessarily change ΔG. Therefore we assume $f^T(E_0) = 10s^{-1}$ independent of E_0 throughout this chapter.

We use two kinds of energy landscape in simulation. The first is the same as in chapter 3, i.e. $E_n = 0$ with a probability of 15/16, and $E_n = \beta$ with a probability of

1/16. The values of k and D are those obtained in chapter 3. In the second kind, the landscape is the same as in the Appendix B, i.e. there are 16 possible values that E_n can take with equal probability, corresponds to all possible bp couples, and $E_n = -4.47$ for GG. The k and D values are those from Appendix B as well.

In both of these cases, we only change the energy of the PAM (E_0 and other PAMs), and investigate how this affect the (average) total searching time. We take for t_{3D} a constant value of $1.39s$. We choose a genome size of $2L+1 = 5001$ (comparable to that of a small bacteriophage genome). The simulation always starts from a 1D diffusion with a uniformly distributed initial condition. For each value of PAM energy, at least 1000 trajectories are simulated, each until it reaches the R-mode. Then we compute the average total searching time for that PAM energy $\langle T_{tot} \rangle$, along with its standard deviation, as its error.

4.1.2 Simulation results

Our simulations reveal a mininum in $\langle T_{tot} \rangle$ as a function of the PAM energy, see fig 4.2. This confirms the presence of a trade-off: If the PAM energy is too low, the Cas9 will waste too much time on PAMs other than the origin, hence increasing $\langle T_{tot} \rangle$. But if the PAM energy is too high, there is a possibility of diffusing away before transit to the R-mode. This may even lead to a further detachment, hence increasing $\langle T_{tot} \rangle$ as well.

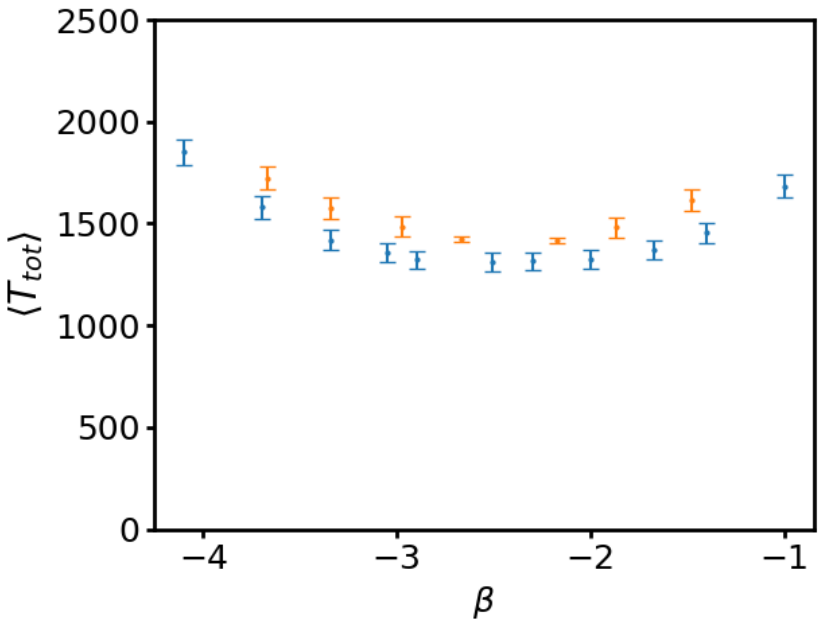


Figure 4.2: $\langle T_{tot} \rangle(s)$ as a function of PAM energy. For the blue data sets, the probability distribution of E_N is the same as that in section 3.2 and for the orange, the bp-dependent model as in Appendix B. The latter model has different energy reference point from the former (as explained in appendix B, the zero energy corresponds to the non-canonical PAM with the highest energy), so orange data points have been shifted horizontally to match their energy reference point.

Near the minimum position, too many trajectories are needed to make the error bar of $\langle T_{tot} \rangle$ small enough. Also longer simulation time is needed for longer genome. To encompass these issues, we attempt the possibility to predict $\langle T_{tot} \rangle$ analytically in the next section.

4.2 Analytical prediction of $\langle T_{tot} \rangle$

We denote the stochastic total search time by T_{tot}. In this chapter, we redefine the average denoted by $\langle ... \rangle$ with one more operation compared with the definition in chapter 2: average over different energy landscape, besides over stochasticity of trajectories and over the initial position n.

4.2.1 $\langle T_0 \rangle$ for a rough and disordered energy landscape

To derive an expression for $\langle T_{tot} \rangle$, we need to first generalize the calculation in section 2.2 to a rough, disordered energy landscape and also discrete lattice. $\langle T_0 \rangle$ is now the average time to reach the origin.

In parallel with the $p(x)$ in chapter 2, we call p_n the probability to find the target within a 1D process, given the process started at position n. We recall that this is a time independent function. In the general case with diffusion and detaching, its recursion relation can be written as

$$p_n = \frac{D_{n+1,n}}{D_{n+1,n} + D_{n-1,n} + k_n} p_{n+1} + \frac{D_{n-1,n}}{D_{n+1,n} + D_{n-1,n} + k_n} p_{n-1}. \tag{4.2}$$

Given our model Eq. (4.1), this equation becomes

$$p_n = \frac{D}{2D + k}(p_{n+1} + p_{n-1}) \tag{4.3}$$

and then

$$p_{n+1} + p_{n-1} - 2p_n = \frac{k}{D} p_n. \tag{4.4}$$

Since $p_0 = 1$ and we assumed the DNA chain is very long, the solution is

$$p_n = \exp\left(-\alpha |x|\right) \tag{4.5}$$

in which $\alpha \equiv \cosh^{-1}\left(\frac{k+2D}{2D}\right)$.

The continuous version of Eq. (4.4) above, assuming $k(x) = k\exp(E(x))$ and $D(x) = D\exp(E(x))$ is

$$\frac{d^2p(x)}{dx^2} = \frac{k}{D}p(x), \tag{4.6}$$

whose solution is

$$p(x) = \exp\left(-\sqrt{\frac{k}{D}}|x|\right). \tag{4.7}$$

The above two equations are the same as Eqs. (2.13) and (2.14) in chapter 2. This is because, for the model defined by Eq. 4.1, Eq. (4.4) and Eq. (4.6) are exactly the same for all n and x, respectively, independent of the distribution of E_n or $E(x)$, since the $\exp(E_n)$ factor cancels out. So the solution is also independent of the energy landscape.

In parallel with $T(x)$, we denote the MFPT (given it started at n) by T_n. This is the mean time to reach the origin rather than the R-mode. The details in its definition are the same as for $T(x)$. It is also a time independent function, and its recursion relation can be written by considering what happens in a time interval δt:

$$T_n = T_n(1 - k_n\delta t - D_{n+1,n}\delta t - D_{n-1,n}\delta t) + \delta t p_n + D_{n+1,n}\delta t T_{n+1} + D_{n-1,n}\delta t T_{n-1}. \tag{4.8}$$

For the model of Eq. (4.1), this simplifies to

$$T_{n+1} + T_{n-1} - 2T_n - \frac{k}{D}T_n + \frac{\exp(E_n)}{D}p_n = 0 \tag{4.9}$$

The continuous version of this equation, using $k(x) = k\exp(E(x))$ and $D(x) = D\exp(E(x))$ is

$$\begin{aligned} &\frac{d^2T(x)}{dx^2} - \frac{k(x)}{D(x)}T(x) + \frac{1}{D(x)}p(x) = \\ &\frac{d^2T(x)}{dx^2} - \frac{k}{D}T(x) + \frac{e^{-E(x)}}{D}p(x) = 0, \end{aligned} \tag{4.10}$$

with the boundary condition $T(0) = 0$, $T(\infty) = 0$.

In the rest of this subsection, the focus will be on the continuous model, since it is easier to solve analytically, and also because we shall see that the result can be applied with very high accuracy to the discrete model.

Apart from the case considered in chapter 2 where $E(x) \equiv 0$, we cannot solve Eq. (4.10) directly, but since what we need is just $\langle T(x) \rangle$, we can proceed as follows.

The average over x and ensemble average (over different energy landscapes) commute, so we first take the ensemble average of Eq. (4.10). Let us denote the ensemble average by $\overline{T}(x)$ (and the same for other quantities). By our definitions,

$\frac{1}{2L}\int_{-L}^{L}\overline{T}(x)dx = \langle T\rangle = \langle p(x)t_{1D}(x)\rangle$, etc. Note that ensemble average and derivatives commute as well, so that we have:

$$\frac{d^2\overline{T}(x)}{dx^2} - \frac{k}{D}\overline{T}(x) + \frac{\overline{\exp(-E(x))}}{D}p(x) = 0. \tag{4.11}$$

As we showed, in this model $p(x)$ is the same for all members of the ensemble, so that $\overline{p(x)} = p(x)$ and $\overline{\exp(-E(x))p(x)} = \overline{\exp(-E(x))}p(x)$. Now, since $\overline{\exp(-E(x))}$ is a constant (independent of x) due to homogeneity, we can solve the above equation in parallel with Eq. (2.18). The solution is just Eq. (2.19) multiplied by a factor of $\overline{\exp(-E(x))}$:

$$\overline{T}(x) = \overline{\exp(-E(x))}\frac{|x|}{2\sqrt{kD}}\exp\left(-\sqrt{\frac{k}{D}}|x|\right). \tag{4.12}$$

We define the mean failed search time T_n^f, in parallel with $T^f(x)$ defined in chapter 2, as the mean time of 1D diffusion contributed by the trajectories that failed to reach the origin before detachment. Here, for a rough energy landscape, the equation for $T^f(x)$ is the same as that of $T(x)$ but with $p(x)$ replaced by $1-p(x)$. The following calculations are also similar. The counterpart of (4.11) is

$$\frac{d^2\overline{T^f}(x)}{dx^2} - \frac{k}{D}\overline{T^f}(x) + \frac{\overline{\exp(-E(x))}}{D}(1-p(x)) = 0. \tag{4.13}$$

and its solution is

$$\overline{T^f}(x) = \overline{\exp(-E(x))}\left(-\frac{|x|}{2\sqrt{kD}}\exp\left(-\sqrt{\frac{k}{D}}|x|\right) + \frac{1}{k}\left[1-\exp\left(-\sqrt{\frac{k}{D}}|x|\right)\right]\right). \tag{4.14}$$

The calculation of $\langle T_0\rangle$ can be done following the same steps as in subsection 2.2.3 until Eq. (2.24):

$$\langle T_0\rangle = \frac{\langle p(x)t_{1D}(x)\rangle}{\langle p(x)\rangle} + \frac{\left\langle (1-p(x))t_{1D}^f(x)\right\rangle}{\langle p(x)\rangle} + \langle t_{3D}\rangle\frac{1-\langle p(x)\rangle}{\langle p(x)\rangle}. \tag{4.15}$$

The average $\langle p(x)\rangle = \frac{1}{2L}\int_{-L}^{L}p(x)dx$ is also unchanged since $p(x)$ stays the same. $\langle p(x)t_{1D}(x)\rangle$ and $\left\langle (1-p(x))t_{1D}^f(x)\right\rangle$ are given by $\frac{1}{2L}\int_{-L}^{L}\overline{T}(x)dx$ and $\frac{1}{2L}\int_{-L}^{L}\overline{T^f}(x)dx$, respectively, So by using Eq. (4.12) and (4.14) we obtain

$$\langle p(x)t_{1D}(x)\rangle = \overline{\exp(-E(x))}\left[-\frac{1}{2k}e^{-\sqrt{\frac{k}{D}}L}+\frac{1}{kL}\sqrt{\frac{k}{D}}(1-e^{-\sqrt{\frac{k}{D}}L})\right] \tag{4.16}$$

and

$$\left\langle (1-p(x))t^f_{1D}(x)\right\rangle = \overline{\exp(-E(x))}\left[\frac{1}{2k}e^{-\sqrt{\frac{k}{D}}L}-\frac{2}{kL}\sqrt{\frac{k}{D}}(1-e^{-\sqrt{\frac{k}{D}}L})+\frac{1}{k}\right]. \tag{4.17}$$

We substitute these three components to find

$$\begin{aligned}\langle T_0\rangle &= \overline{\exp(-E(x))}\left(\frac{L}{\sqrt{kD}(1-e^{-\sqrt{\frac{k}{D}}L})}-\frac{1}{k}\right)+\langle t_{3D}\rangle\left(\sqrt{\frac{k}{D}}\frac{L}{1-e^{-\sqrt{\frac{k}{D}}L}}-1\right)\\ &\approx \overline{\exp(-E(x))}\frac{L}{\sqrt{kD}}+\langle t_{3D}\rangle L\sqrt{\frac{k}{D}}.\end{aligned} \tag{4.18}$$

Comparing this equation with Eq. (2.25), we see that under the model of Eq. (4.1), the 1D sliding time is scaled by a factor of $\overline{\exp(-E(x))}$, which is intuitively clear, because both diffusion and detachment on a position with energy $E(x)$ are scaled by a factor $\overline{\exp(-E(x))}$.

Finally, we comment on the result. In facilitated diffusion, the mean searching time is often estimated as

$$\langle T_0\rangle \approx \Gamma(\langle t_{1D}\rangle + \langle t_{3D}\rangle) \tag{4.19}$$

(see for example, [Cencini and Pigolotti, 2018]), in which Γ is the number of 1D and 3D diffusion rounds. $\frac{1}{\langle p(x)\rangle}$ is by definition the average number of rounds, so all three terms in (4.15) have the factor $\frac{1}{\langle p(x)\rangle}$. The other insight is that, from Eq. (4.7) we know that $\langle p(x)\rangle$, which is just the mean of $p(x)$ over $(-L, L)$ (essentially over $(-\infty, +\infty)$), is mainly contributed by values of $p(x)$ where $|x|$ is small. This is the message conveyed in [Cencini and Pigolotti, 2018]: the number of rounds is determined by how easy it is to find the target when sliding in its proximity, rather than by the average sliding length.

The $1-\langle p(x)\rangle$ factor in the t_{3D} term appears because the search process starts with a 1D diffusion, so $\langle t_{3D}\rangle$ should be multiplied by "the number of rounds minus one", i.e. $\frac{1}{\langle p(x)\rangle}-1$ rather than just $\frac{1}{\langle p(x)\rangle}$.

4.2.2 The framework of the calculation

The time T_{tot} is distributed according to a certain distribution $P(T_{tot})$, so that $\int_0^{+\infty} T_{tot}P(T_{tot})dT_{tot} = \langle T_{tot}\rangle$. In the calculation of $\langle T_{tot}\rangle$ we do not need to compute $P(T_{tot})$ explicitly: what we only use is its normalization, $\int_0^{+\infty} P(T_{tot})dT_{tot} = 1$.

The time T_0 is also a stochastic quantity with a distribution $P_0(T_0)$, such that $\int_0^{+\infty} T_0P_0(T_0)dT_0 = \langle T_0\rangle$. In the calculation of $\langle T_{tot}\rangle$, we only use $\int_0^{+\infty} P_0(T_0)dT_0 = 1$ as well.

To start our derivation, we first divide the trajectories into two types:

- In the first type, between the first binding to the central PAM and the final recognition, there is no detachment and 3D diffusion.
- In the second type, there are one or more detachment events.

Now, we make the approximation that after Cas9 reaches the central PAM, there are only two possibilities for its next step: it either transits into the R-mode, or it diffuse to one of the central PAM's two neighbours. In other words, we neglect the probability of directly detaching from the central PAM. This is a good approximation because $ke^{\beta} \ll 2De^{\beta} \lesssim f^T$, and it substantially simplifies the calculation. The consequence of this approximation is that a Cas9 at the central PAM has probability $\frac{f^T}{f^T+2De^{\beta}}$ to transit into the R-mode, and probability $\frac{2De^{\beta}}{f^T+2De^{\beta}}$ to diffuse away. We denote these probabilities by p and $1-p$, respectively.

The contributions from trajectories in the two types can be calculated by dividing them further into different groups. The detailed calculation is given in appendix E. Here we just express the result. The contribution to $\langle T_{tot}\rangle$ from the first type of trajectories is

$$p\langle T_0\rangle\frac{1}{1-r} + \frac{p(1-p)e^{-\sqrt{\frac{k}{D}}}}{f^T+2De^{\beta}}\frac{1}{(1-r)^2} + \frac{p(1-p)e^{-\sqrt{\frac{k}{D}}}}{\sqrt{4Dk}}\frac{1}{(1-r)^2} + \frac{p}{f^T+2De^{\beta}}\frac{1}{1-r}, \tag{4.20}$$

in which $r = \left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)$. The contribution from the second type is

$$\begin{aligned}&(\langle T_0\rangle + t_{3D} + \langle T_{tot}\rangle)(1-e^{-\sqrt{\frac{k}{D}}})(1-p)\frac{1}{1-r}\\ +\frac{(1-p)}{f^T+2De^{\beta}}(1-e^{-\sqrt{\frac{k}{D}}})\frac{1}{(1-r)^2} &+ \frac{(1-p)^2e^{-\sqrt{\frac{k}{D}}}}{\sqrt{4Dk}}(1-e^{-\sqrt{\frac{k}{D}}})\frac{1}{(1-r)^2}\\ &+\frac{1}{k}\left(1-e^{-\sqrt{\frac{k}{D}}}\left(1+\sqrt{\frac{k}{4D}}\right)\right)(1-p)\frac{1}{1-r}.\end{aligned} \tag{4.21}$$

4.2.3 $\langle T_{tot}\rangle$ solved

Finally, by equating $\langle T_{tot}\rangle$ with the sum of Eqs. (4.20) and (4.21), we can solve $\langle T_{tot}\rangle$ as an unknown (note that it also appears in Eq. (4.21), see the derivation in Appendix

E for the reason).

$$\begin{aligned}\langle T_{tot}\rangle = \langle T_0\rangle + \frac{(1-p)e^{-\sqrt{\frac{k}{D}}}}{f^T+2De^{\beta}}\frac{1}{(1-r)} + \frac{(1-p)e^{-\sqrt{\frac{k}{D}}}}{\sqrt{4Dk}}\frac{1}{(1-r)} + \frac{1}{f^T+2De^{\beta}} \\ +(\langle T_0\rangle + t_{3D})(1-e^{-\sqrt{\frac{k}{D}}})\frac{(1-p)}{p} + \frac{(1-p)}{f^T+2De^{\beta}}(1-e^{-\sqrt{\frac{k}{D}}})\frac{1}{(1-r)p} \\ +\frac{(1-p)^2e^{-\sqrt{\frac{k}{D}}}}{\sqrt{4Dk}}(1-e^{-\sqrt{\frac{k}{D}}})\frac{1}{(1-r)p} + \frac{1}{kp}\left(1-e^{-\sqrt{\frac{k}{D}}}\left(1+\sqrt{\frac{k}{4D}}\right)\right)(1-p)\end{aligned} \tag{4.22}$$

By substitute Eq. (4.18), and since $L \gg 1$, this can be considerably simplified into

$$\langle T_{tot}\rangle \approx L\left[1+\frac{2De^{\beta}}{f^T}(1-e^{-\sqrt{\frac{k}{D}}})\right]\left(\frac{\overline{\exp(-E_n)}}{\sqrt{kD}} + t_{3D}\sqrt{\frac{k}{D}}\right) \tag{4.23}$$

By taking the derivative of the above equation, we can find the β correspond to the minimum $\langle T_{tot}\rangle$. One can see from Eq. 4.23 that the result does not depend on L. The result is

$$\beta \approx -\frac{1}{2}\ln\left[\frac{2D}{\alpha f^T}(1-e^{-\sqrt{\frac{k}{D}}})\left((1-\alpha)\overline{\exp(-E_n)}' + t_{3D}k\right)\right], \tag{4.24}$$

in which α is the abundance of PAM (e.g. 1/16 for a two-bp PAM), and $\overline{\exp(-E_n)}'$ is the average of $\exp(-E_n)$ for n that is not a PAM. In formula,

$$(1-\alpha)\overline{\exp(-E_n)}' + \alpha e^{\beta} = \overline{\exp(-E_n)}. \tag{4.25}$$

Then we analyse the dependence of the optimal β on our model parameters as follows. Varying one parameter on the RHS of Eq. (4.24) at a time, while keeping the others constant, we find that:

1, The optimal β increases with increasing α. This is because the more abundant the PAMs are, the more time will be wasted on irrelevant PAMs. A larger β can compensate for this and hence leads to a smaller $\langle T_{tot}\rangle$.

2, The optimal β decreases with increasing k. With a higher k, the cost of missing the target is higher: the risk of detaching and doing another 1D round is larger. A smaller β can compensate for this and hence leads to a smaller $\langle T_{tot}\rangle$.

3, The optimal β decreases with increasing t_{3D}. The reason is similar to when k is increasing: the cost of missing the target is higher due to the larger t_{3D}.

4, The optimal β decreases with increasing $\overline{\exp(-E_n)}'$. Since the 1D diffusion time spent on non-PAM sites is proportional to $\overline{\exp(-E_n)}'$, The reason is again similar to the last two cases: the cost of missing the target is higher with a higher $\overline{\exp(-E_n)}'$.

5, The dependence of the optimal β on D is not monotonic. This is because D

affects both sides of the trade-off.

In summary, the two approximations used in this section are:

1, We ignored the probability that Cas9 detaches directly from $n = 0$. This approximation neglects some detaching events. This is the possible reason to the fact that, when t_{3D} is small, the precision of the result is higher, since the proportion of 3D diffusion time in $\langle T_{tot} \rangle$ is smaller when t_{3D} is small.

2, We used the first passage distribution of a 1D Brownian motion (rather than that of a 1D random walk in discrete space, see appendix E).

If t_{3D} is taken to be a stochastic quantity, the only change in the calculation is to replace the constant t_{3D} by $\langle t_{3D} \rangle$.

4.2.4 Comparison with simulation and predictions

The analytical prediction for the two-level landscape compared with simulations is shown in fig. 4.3. For the worst data point, the analytical result is still within two the standard errors from the simulations.

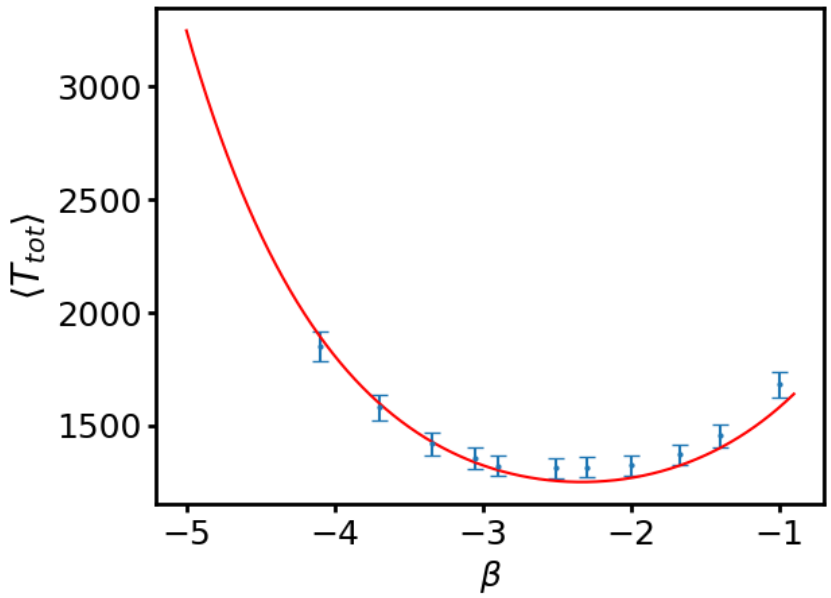


Figure 4.3: Comparison of analytical result and simulation. Model choice and parameters are the same as in fig. 4.2.

For smaller t_{3D} value, the prediction is even better. Fig. 4.4 shows the result for $t_{3D} = 0.139s$, which is 1/10 of that in fig. 4.3. Other parameters are the same.

For longer genome, simulations become slow. For example, for a 10^4 bp genome, result for $\beta = -2.5$ and -2.3 are $\langle T_{tot} \rangle = 2543.1s$ and $\langle T_{tot} \rangle = 2543.4s$, respectively. 10000 trajectories result in an errorbar of around 25 s, therefore 10^8 trajectories are not enough to locate the position of the mininum precisely.

Fig 4.5 shows the analytical results for genome sizes 5000, 10000, 15000, 20000. The other configurations and parameters the same as in fig 4.3. This demonstrate that

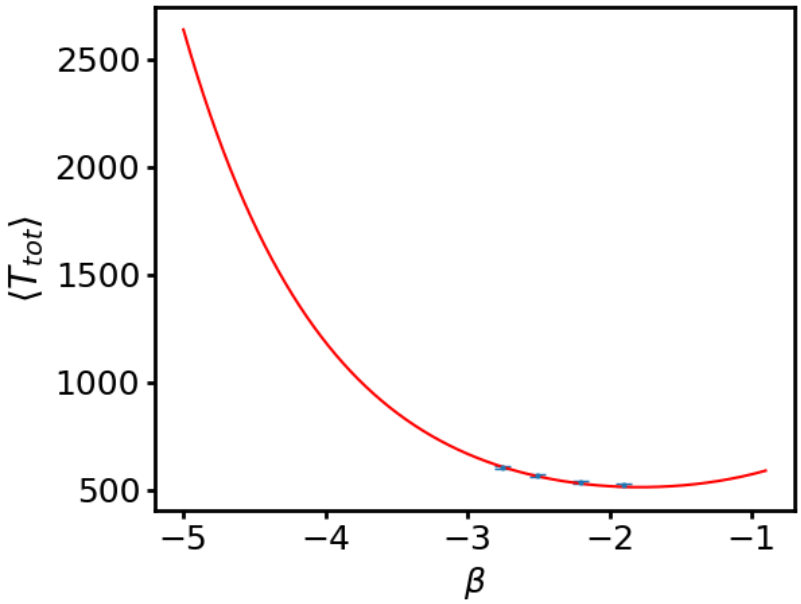


Figure 4.4: Comparison of analytical result and simulation for $t_{3D} = 0.139$. Model choice and other parameters are the same as in fig 4.3

the prediction of Eq. (4.23) is indistinguishable from that of Eq. (4.22), and hence Eq. (4.24) is very precise, too.

The result shows that for large genome, this trade-off is very significant. Since bacteriophage genome ranges from 5,000 to 5,000,000 bp, a typical genome is likely to be even larger than those considered in fig. 4.5, so the mean total search time $\langle T_{tot} \rangle$ could be very sensitive to the PAM energy.

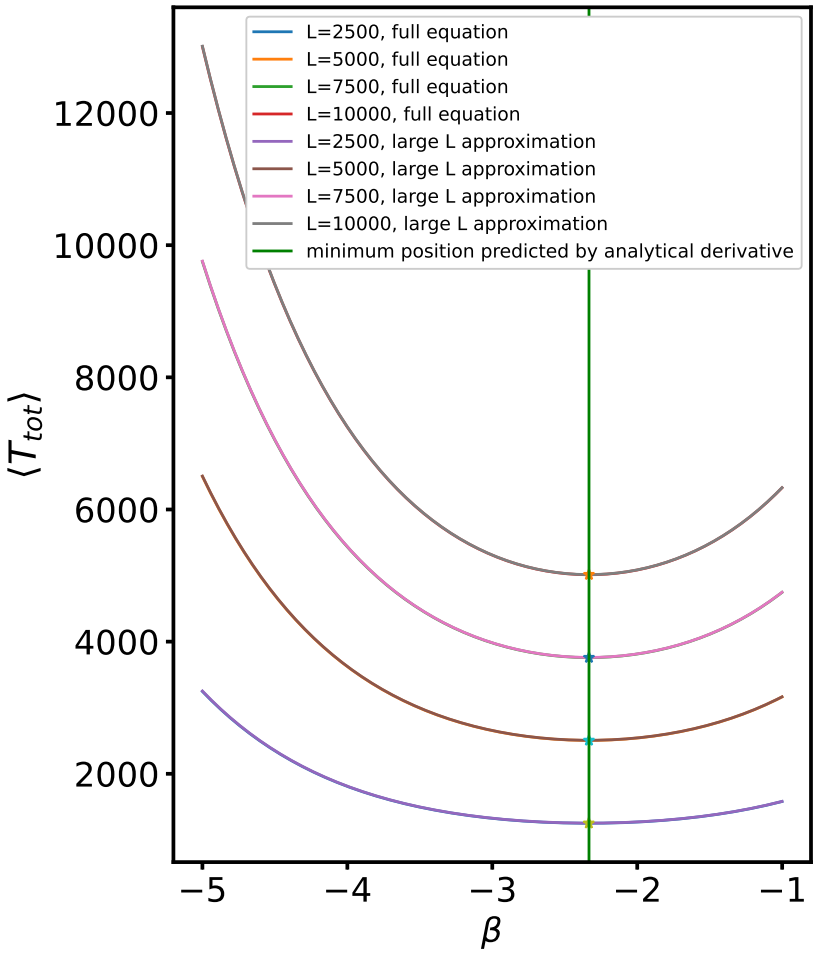


Figure 4.5: Analytical results of $\langle T_{tot} \rangle$, for genome sizes 5000, 10000, 15000, 20000. Other parameters the same as in fig 4.3. The minimum position is marked. One can only see four curves rather than eight, because the four curves by Eq. (4.23) are indistinguishable from their four counterparts by Eq. (4.22).

Chapter 5

Specificity of Cas9 recognition of its target

At variance with the previous two chapters, the results in this chapter are only preliminary. We study the specificity of Cas9 recognition of its main target. The first section is based on the measured equilibrium occupancy of Cas9 on targets with mismatches. We attempt to use these occupancy data to infer nucleic acid thermodynamics parameters, and predict the binding free energy of other unmeasured mismatched targets. In the second section, we propose a model to explain that the association/disassociation rates are observed to be constant in time. This fact has been introduced in subsection 1.1.3. Both sections of this chapter are based on the experimental results from [Boyle et al., 2017].

5.1 The specificity energetics

As explained in the introduction, [Boyle et al., 2017] creates variants of a given DNA target sequence, where 1 or 2 bp of the 20 bp main target are mutated. They measured equilibrated occupancies of the variants after incubation. Like other experiment focusing on off-target behaviour, dCas9 instead of Cas9 was used. From now on, we call a target sequence with 1/2 mutated bp as "single mismatch"/"double mismatch". In this section, we present a model that uses these results to infer nucleic acid thermodynamics parameters, and predict the binding free energy and occupancy of other unmeasured mismatched targets that contains more than two mutated base pairs.

5.1.1 The model without compensation terms

In [Boyle et al., 2017], the occupancy was measured in such a way that only long-time binding of Cas9 was recorded. Also, the target DNA sequences were so short that every sequence can either be occupied by only one (rather than multiple) Cas9 or empty. Therefore, we do not have to consider any binding other than binding to the PAM next to the main target or to the target itself. We hence assume that, for a DNA sequence in the experiment of [Boyle et al., 2017], there are 22 states. The first of them is the empty unbound state, the second being the PAM state, and the rest 20 states

corresponds to the sequential pairings of the target strand with gRNA.

The energy of the PAM state is given in chapter 3. We denote its value by ϵ_{PAM}, while the unbound state has energy 0. A state i among the 20 states (i from 1 to 20) has energy ϵ_i. For a dsDNA or a DNA-RNA heteroduplex, the Gibbs free energy depends on nearest neighbour (NN) bp pairs [SantaLucia Jr, 1998], so we have

$$\epsilon_{i+1} - \epsilon_i = -\Delta G\begin{pmatrix} d_{i+2}d_{i+1} \\ d'_{i+2}d'_{i+1} \end{pmatrix} + \Delta G\begin{pmatrix} d'_{i+1}d'_i \\ r_{i+1}r_i \end{pmatrix}. \tag{5.1}$$

Here, d_i and d'_i represent the bases at position i in the non-target and target DNA strands, respectively. r_i represents the bases at position i in the gRNA. Each of the parentheses corrsponds to a nearest neighbour (NN) bp pair. The terms $\Delta G(..)$ are the Gibbs free energies of forming such NN pairs. The sign convention is such that ΔG is more negative for stronger matches. Equation (5.1) expresses the free energy difference $\epsilon_{i+1} - \epsilon_i$ in terms of the energy gain when "flipping" a base (d'_{i+1}) from the double-strand DNA side to the DNA-RNA hybrid side. In general, this energy gain can be positive or negative if there is no mismatch in the hybrid NN pair, but is likely to be positive if there is a mismatch. We do not assume an ad hoc compensation energy term coming from the interaction with Cas9 at this stage. During the hybrid-forming process, Cas9 might facilitate the unwinding. However, whether the hybrid can form successfully mainly depends on the matching between the gRNA and the DNA. In other words, Cas9 can only read its target by gRNA, but not directly as a TF would do. Besides, Cas9 does not consume ATP. Based on the above two facts, we assume that the energetics is mainly determined by the specific base pair sequences, rather than the intervene of Cas9. The above analysis also implies that, even if there is some facilitation by Cas9, we can assume it to be independent of the specific base pairs. This possibility is considered in the other model reviewed in section 5.1.4.

The energy for state i is hence

$$\epsilon_i = \epsilon_{PAM} + \epsilon_1 + \sum_2^i \left(-\Delta G\begin{pmatrix} d_{i+1}d_i \\ d'_{i+1}d'_i \end{pmatrix} + \Delta G\begin{pmatrix} d'_i d'_{i-1} \\ r_i r_{i-1} \end{pmatrix} \right). \tag{5.2}$$

The energy ϵ_1 is treated as an independent parameter. The grandcanonical partition function for a particular DNA sequence x in the occupancy experiment reads

$$Z_x = 1 + e^{-(\epsilon_{PAM}-\mu)} + \sum_1^{20} e^{-(\epsilon_i-\mu)}, \tag{5.3}$$

where the first term "1" corresponds to the unbound state and μ is the chemical potential. The chemical potential μ is also treated as a independent parameter, and is the same for all target DNA sequences, since in the experiment they are all in the same buffer. With a different target DNA sequence (but the same gRNA), the partition function would be different. The partition function can alternatively be written as

$$Z_x = 1 + e^{-(\epsilon_{bind}-\mu)}, \tag{5.4}$$

in which $\epsilon_{bind} = e^{-\epsilon_{PAM}} + \sum_1^{20} e^{-\epsilon_i}$.

Experiments show that a mismatch near the PAM is more relevant than a mismatch far away. This can emerge from our model, because a mismatch at position i increases the energy of all states after i, so that earlier mismatches have a larger influence than later ones.

In the literature, the nucleic acid thermodynamic parameters of double strand DNA NN pairs, as well as (mis)matched DNA-RNA NN pairs are usually measured in 1 M NaCl condition, which is not the case in [Boyle et al., 2017]. Therefore, we can use this model to predict all the nucleic acid thermodynamic parameters in the ionic condition of [Boyle et al., 2017].

In this model, there are 122 parameters and 1599 data points. We shall divide these data into a training set and a test set.

5.1.2 The likelihood function of the model by Bayesian approach

The ultimate goal is to maximize the posteriori probability of our model M: $prob(M \mid data, I)$, where I is all the background information except the experimental data at hand. We express this probability as

$$prob(M \mid data, I) \propto prob(data \mid M, I) \times prob(M \mid I). \tag{5.5}$$

The proportional factor is equal to $1/prob(data \mid I)$. We assume a uniform probability $prob(M \mid I)$, so we will focus on $prob(data \mid M, I)$ from now on.

If a specific target DNA sequence x has N_x copies (we will adopt the name "clusters" as in [Boyle et al., 2017]) and the observed occupancy is O_x, then we assume that $N_x O_x$ of them are bound, and $N_x(1 - O_x)$ of them are unbound.

The occupancy predicted by the model for sequence x is $\hat{O}_x = 1 - 1/Z_x$. In the experiment, the bound/unbound state of a cluster should be independent of that of other clusters. Then, the probability that we observe the data for a given sequence x given our model is

$$prob(N_x, O_x \mid \hat{O}_x, I) \propto \hat{O}_x^{N_x O_x}(1 - \hat{O}_x)^{N_x(1-O_x)}, \tag{5.6}$$

where we omitted the combinatorial prefactors, since they only depend on N_x and O_x, but not on our model nor its parameters. As said, the state of bound/unbound of every cluster should be independent of other clusters. This also holds between different sequences (i.e. different x). So we immediately arrive at

$$\mathcal{L} = prob(data \mid M, I) \propto \prod_x \hat{O}_x^{N_x O_x}(1 - \hat{O}_x)^{N_x(1-O_x)}. \tag{5.7}$$

This is the likelihood to be maximized. In practice, we minimize minus the log likelihood

$$
\begin{aligned}
-\ln \mathcal{L} = -\sum_x \Big(N_x O_x \ln \hat{O}_x + N_x (1 - O_x) \ln(1 - \hat{O}_x) \Big) = \\
-\sum_x N_x \Big(O_x \ln \hat{O}_x + (1 - O_x) \ln(1 - \hat{O}_x) \Big) .
\end{aligned}
\tag{5.8}
$$

In terms of our partition functions, this quantity is expressed by

$$
-\ln \mathcal{L} = -\sum_x N_x \left(O_x \ln(1 - \frac{1}{Z_x}) - (1 - O_x) \ln Z_x \right) . \tag{5.9}
$$

5.1.3 Approximation by the central limit theorem

We now derive a approximation of Eq. (5.9) by using the central limit theorem, which will facilitate the numerical minimization. Since N_x is usually large (around 1000 clusters for single mismatches and 100 for double mismatches), the binomial distribution in Eq. (5.6) can be approximated by a Gaussian by the central limit theorem:

$$
prob(N_x, O_x \mid \hat{O}_x, I) = \binom{N_x}{N_x O_x} \hat{O}_x^{N_x O_x} (1 - \hat{O}_x)^{N_x (1 - O_x)} \propto \exp\left(-\frac{(N_x O_x - N_x \hat{O}_x)^2}{2 N_x \hat{O}_x (1 - \hat{O}_x)} \right), \tag{5.10}
$$

To show this, first notice that the standard De Moivre-Laplace Theorem takes the form

$$
\binom{n}{k} p^k q^{n-k} \to \frac{1}{\sqrt{2\pi npq}} \exp - \frac{(k - np)^2}{2npq}. \tag{5.11}
$$

In our problem, N_x, O_x, $\hat{O}_x$ and $1 - \hat{O}_x$ amounts to n, nk, p and q, respectively. By comparing the LHS of Eq. (5.10) and that of (5.11), one may think that on the RHS of Eq. (5.10) the proportionality factor is $\frac{1}{\sqrt{2\pi N_x \hat{O}_x (1 - \hat{O}_x)}}$, which depends on $\hat{O}_x$, so cannot be omitted. But in fact, one of the step in the standard proof of Eq. (5.11) is exactly to replace $\frac{1}{\sqrt{2\pi n \frac{k}{n}(1 - \frac{k}{n})}}$ by the final $\frac{1}{\sqrt{2\pi npq}}$, by using $\frac{k}{n} \to p$. So, we can simply undo that step, and our hidden prefactor in the RHS of (5.10) is $\frac{1}{\sqrt{2\pi N_x O_x (1 - O_x)}}$, which depends on O_x rather than $\hat{O}_x$, so can be omitted and does not affect the likelihood. All other steps are the same as the standard proof of Eq. (5.11).

Then, we express the log-likelihood in a weighted sum of square form:

$$
-\ln \mathcal{L} = \sum_x \frac{N_x}{\hat{O}_x (1 - \hat{O}_x)} (O_x - \hat{O}_x)^2. \tag{5.12}
$$

This amounts to a sum of squares weighted by their error bar $\sqrt{\frac{N_x}{\hat{O}_x (1 - \hat{O}_x)}}$. The advantage of using the central limit theorem and using Eq. (5.12) is that the minimization of the sum of squares is much more efficient than most of other function forms.

5.1.4 The model with compensation terms

We can also add compensation terms E_{ci}, representing the energy gain due to the interaction of DNA/gRNA with Cas9. We assume that these energies do not depend on the bp sequences. So that Eq. (5.1) becomes

$$\epsilon_{i+1} - \epsilon_i = -\Delta G\begin{pmatrix} d_{i+2}d_{i+1} \\ d'_{i+2}d'_{i+1} \end{pmatrix} + \Delta G\begin{pmatrix} d'_{i+1}d'_i \\ r_{i+1}r_i \end{pmatrix} + E_{ci}, \tag{5.13}$$

for E_{ci}, i can take integers from 1 to 20. We note that E_{c1} and ϵ_1 are not independent variables. One can only determine $E_{c1} + \epsilon_1$: in subsection 5.1.1, ϵ_1 is also a parameter that does not depend on the bp sequences, see Eq. (5.2). Hence there are 19 more free parameters than in the model without compensation terms. Equations from (5.2) to (5.4) in the model change accordingly.

5.1.5 Results

By randomly choosing half of the data as the training set, we optimize parameters using Eq. (5.12). Then, the rest of the data are compared with their predicted values from both the model with and without compensation terms using the optimal parameters. The results are shown in fig 5.1.

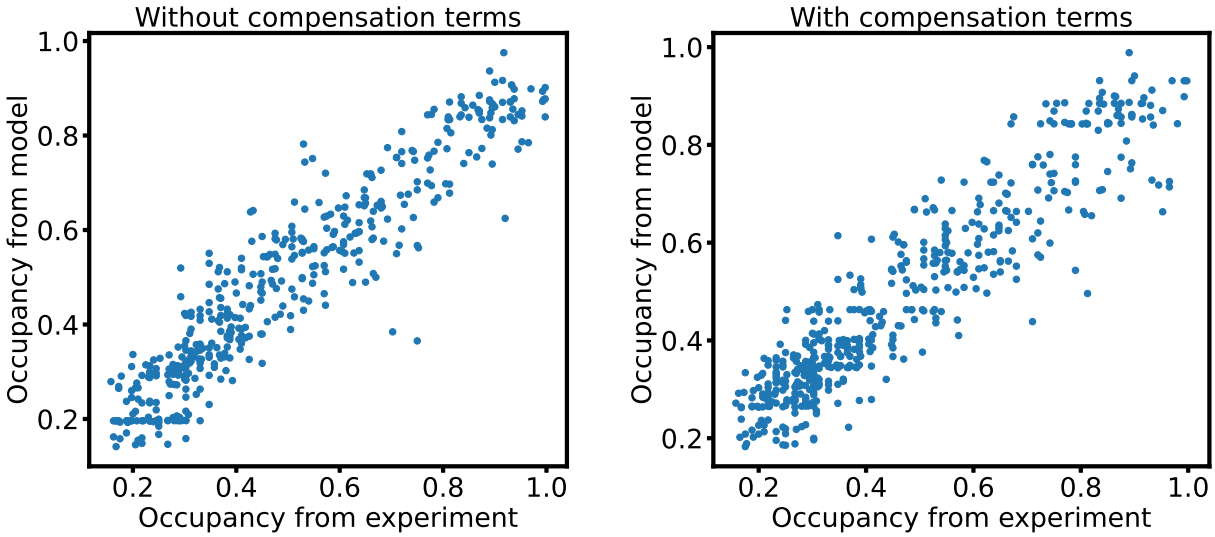


Figure 5.1: Half of the (randomly chosen) occupancy data in test set compared with their predicted values by the model without/with compensation.

We call a random separation of the data into training set and test set as a randomization. Multiple randomizations generate different optimal parameters. We then take the average of the resulting parameters to obtain a final result. Fig 5.2 is the final result in the model without compensation, from the average of 5 randomizations.

A few inconsistencies in the mismatched parameters emerge because, some mismatches appear only a few times in the data set (while matched parameters always appear many times), so their estimation depends on the choice of the training set.

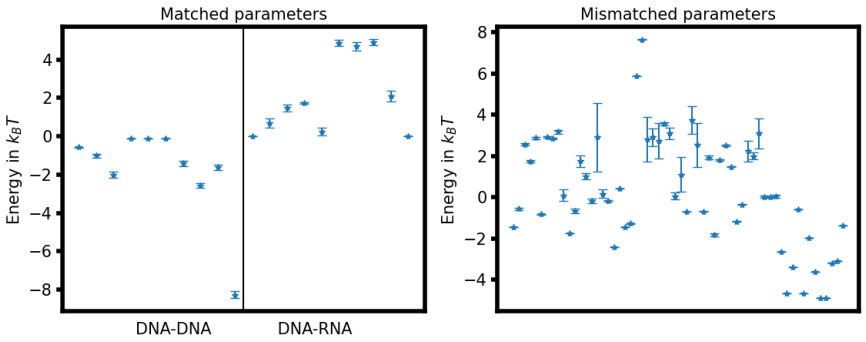


Figure 5.2: The final result in the model without compensation from 5 randomizations. In both panels, the horizontal axis corresponds to different NN pairs but will be too packed if specified. In the left panel, the first 10 are DNA-DNA parameters. Last 10 are DNA-RNA ones (with a minus sign). Error bars are standard deviations.

The corresponding result for the model with compensation is shown in fig 5.3.

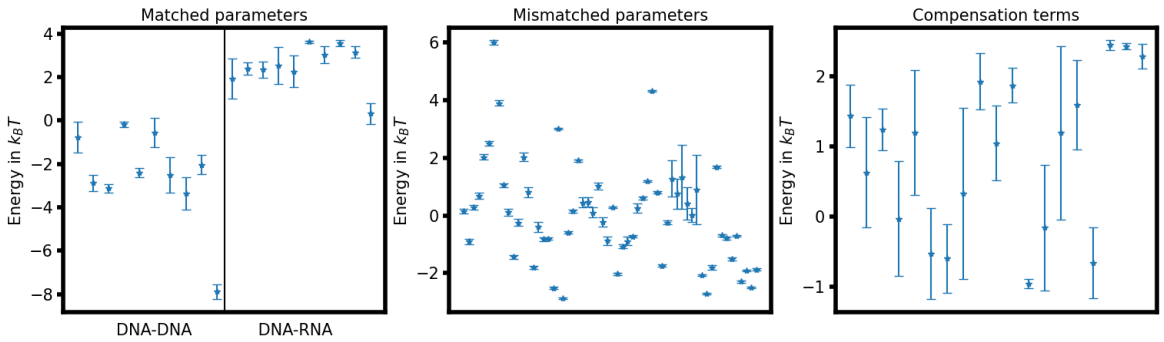


Figure 5.3: Results for the model with compensation by 5 randomizations. In the left panel, the first 10 are DNA-DNA parameters. Last 10 are DNA-RNA ones (with a minus sign). Error bars are standard deviation. In the right panel, the E_{ci} are plotted in the order of their subindices.

In this case the mismatched parameters are more consistent than in the model without compensation. The compensation terms themselves have large error bars, but we note that the vertical axis is much finer than the first two panels.

A comparison of the optimized DNA-DNA NN pairs energy parameters with the literature value is shown in fig 5.4.

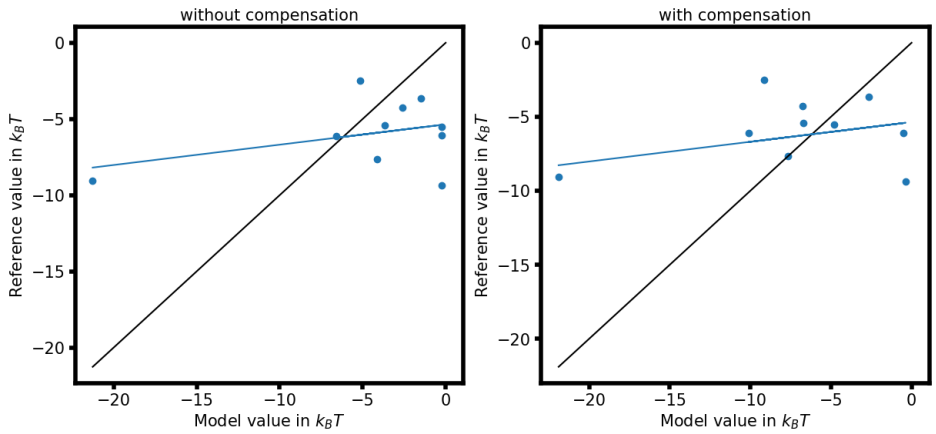


Figure 5.4: The optimized DNA-DNA NN pairs energy parameters (horizontal) compared with the literature value (vertical). Black straight lines are the line $y = x$ and blue straight lines are linear regression results.

The sample Pearson correlation coefficients of the model without/with compensation are $r = 0.374$ and $r = 0.23$, respectively. As pointed out, these literature values are measured in 1 M NaCl condition, but this is not the case in [Boyle et al., 2017]. This might be the cause for the inconsistency.

5.2 The implications of the constant association and disassociation rate

As mentioned in the introduction, the association and disassociation rates measured in [Boyle et al., 2017] are nearly constant for a long time interval (more than 1500 seconds) for all sequences except for those with a negligible rate. Here we still consider a model with 22 states as in last section, but we focus on the transition rates rather than energy. Our idea is that the constancy of rates should impose some constraints on the transition rates. We present a model with a master equation for the 22 states. We will see that the constant rate fact implies that there is a "gap" in the eigenvalue spectrum of the master equation. It may be possible to use a perturbation-expansion-type calculation to extract the implications of the constant-rate fact, but this calculation is relatively difficult and has not been done yet. We here present the model as an open problem.

5.2.1 Association

The 22 states considered here are the same as in last section, but we adopt a different notation. Here, the unbound state has lower index 1, PAM state 2, etc. The transition is between nearest neighbour only, therefore the master equation matrix $\mathbf{A}$ is tridiagonal and 22 by 22. Since initially all DNA sequences are unbound, the initial condition is the vector $\mathbf{p}(0) = (1, 0, ..., 0)$.

The Perron-Frobenius theorem dictates that, as long as all entries of $\mathbf{A}$ are nonzero, the matrix has an unique eigenstate $\boldsymbol{\phi}^{(1)}$ corresponding to the equilibrium distribution

with eigenvalue 0, and all other eigenstates $\boldsymbol{\phi}^{(i)}, i \neq 1$ have real and negative eigenvalue $-\lambda_i$. We order them from small to large magnitude. We expand our solution $\mathbf{p}(t)$ as

$$\mathbf{p}(t) = \boldsymbol{\phi}^{(1)} + \sum_{i=2}^{22} a_i e^{-\lambda_i t} \boldsymbol{\phi}^{(i)} \tag{5.14}$$

in which the $\boldsymbol{\phi}^{(1)}$ is normalized and all other eigenstates satisfy $\sum_j \boldsymbol{\phi}_j^{(i)} = 0, i \neq 1$. a_i are coefficients such that $\mathbf{p}(0) = (1, 0, ..., 0) = \boldsymbol{\phi}^{(1)} + \sum_{i=2}^{22} a_i \boldsymbol{\phi}^{(i)}$.

The fact that the association rate $-\frac{d\mathbf{p}_1}{dt}$ being a constant during the time window of 0 to 2000 seconds implies a "gap" in the eigenvalue spectrum. In other words, if we define a time scale $\tau_1 = 20000s$ which is ten times the time window, some eigenvalues $\lambda_2 ... \lambda_k$ are smaller than $1/\tau_1$, and others $\lambda_{k+1} ... \lambda_{22}$ are much larger than $1/\tau_1$. Then, a linear expansion is justified in the time window 0 to 2000 seconds:

$$\mathbf{p}(t) \approx \boldsymbol{\phi}^{(1)} + \sum_{j=2}^{k} a_j (1 - \lambda_j t) \boldsymbol{\phi}^{(j)} \tag{5.15}$$

and the association rate $-\frac{d\mathbf{p}_1}{dt}$ during this period is (by taking derivative of the above equation) $\sum_{j=2}^{k} a_j \lambda_j \boldsymbol{\phi}_1^{(j)}$, which is a constant of time. If there is no "gap" in the spectrum, then Eq. (5.15) will not hold, and there will not be a time window such that $-\frac{d\mathbf{p}_1}{dt}$ is constant in time, in contradiction to the constant rate fact.

5.2.2 Disassociation

In the disassociation experiment, any detached Cas9 will be washed away at once, so the unbound state is a absorbing state. Therefore the master equation matrix $\mathbf{B}$ has $\mathbf{B}_{11} = \mathbf{B}_{21} = 0$, so the first column is a zero vector. Besides these 2 entries, $\mathbf{B}$ is the same as $\mathbf{A}$. Because the experiment starts with a equilibrium state but with the unbound Cas9 removed, here the initial condition is the vector $\mathbf{q}(0) = \frac{1}{\sum_{i=2^2 2} \boldsymbol{\phi}_i^{(1)}} \left(\boldsymbol{\phi}^{(1)} - (\boldsymbol{\phi}_1^{(1)}, 0, ..., 0) \right)$ (essentially the normalized version of $\boldsymbol{\phi}^{(1)} - (\boldsymbol{\phi}_1^{(1)}, 0, ..., 0)$) .

We know that $\mathbf{B}$ has one equilibrium distribution $\boldsymbol{\psi}^{(1)} = (1, 0, ..., 0)$ with eigenvalue 0, and all other eigenstates $\boldsymbol{\psi}^{(i)}, i \neq 1$ have real and negative eigenvalue $-\mu_i$, and they are ordered from small to large magnitude. We expand our solution $\mathbf{q}(t)$ as

$$\mathbf{q}(t) = \boldsymbol{\psi}^{(1)} + \sum_{i=2}^{22} b_i e^{-\mu_i t} \boldsymbol{\psi}^{(i)} \tag{5.16}$$

in which the $\boldsymbol{\mu}^{(1)}$ is normalized and all other eigenstates satisfy $\sum_j \boldsymbol{\mu}_j^{(i)} = 0, i \neq 1$. b_i are coefficients such that $\mathbf{q}(0) = \boldsymbol{\psi}^{(1)} + \sum_{i=2}^{22} b_i \boldsymbol{\psi}^{(i)}$.

Following the same reasoning as in the last subsection, again we define $\tau_2 = 20000s$, then we have some eigenvalues $\mu_2 ... \mu_k$ that are smaller than $1/\tau_2$, and some $\mu_{k+1} ... \mu_{22}$ that are much larger than $1/\tau_2$, so the linear expansion below is justified in the time window of 0 to 2000 seconds:

$$\mathbf{q}(t) \approx \boldsymbol{\psi}^{(1)} + \sum_{j=2}^{k} b_j(1-\mu_j t)\boldsymbol{\psi}^{(j)}. \tag{5.17}$$

By taking derivative of the above equation, the disassociation rate is $\frac{d\mathbf{q}_1}{dt} - \sum_{j=2}^{k} b_j \mu_j \boldsymbol{\psi}_1^{(j)}$.

A possible way to proceed is as follows. For the on-target sequence, the master equation may be estimated by a master equation corresponds to a biased random walk, but perturbed (especially the rates related to $i = 1$ and $i = 22$). Then, mismatches can be mapped to further perturbations of the system with some "defects" at the states corresponding to mismatched bp.

Conclusion

To summarize, we first studied the target-searching of Cas9 by facilitated diffusion models, in which the PAM plays a important role. The result shows that unlike other DNA-binding proteins such as TF, Cas9 has rather short sliding length. Then we found that the generic distribution of Cas9 on DNA imply a hopping mechanism apart from 1D diffusion along the DNA, and the physics has a deep relationship with the theory of Anderson localization.

We also studied the efficiency of Cas9 searching, focusing on the average time consumed to reach the main target. We found that this average time depends on the energy of PAM, and there is a optimized PAM energy that minimizes it. This is essentially because there is a trade-off between the time spent on irrelevant PAMs and the possibility of missing the main target.

The link with Anderson localization that we formalized may be applied to other proteins that perform facilitated diffusion such as TF. The binding profiles of TF along the DNA can be measured at base pair resolution By using modern immunoprecipitation techniques [Rhee and Pugh, 2011]. However, the interpretation of these binding profiles is still under debate [MacQuarrie et al., 2011]. The sequence-dependent models of facilitated diffusion by TF [Slutsky and Mirny, 2004, Bauer et al., 2015] [Cencini and Pigolotti, 2018] can be combined with the Anderson localization approach. This can lead to more insight on this crucial problem in biophysics.

We then studied the average search time for Cas9 to recognize its target, and found that this time has a minimum as the PAM energy varies. We did a successful analytical calculation that gives the average search time as a function of the PAM energy, as well as of other parameters in the facilitated diffusion model. We can also predict the exact value of the optimal PAM energy that leads to the minimum of this time. These results are in very good agreement with our simulation. As expected, there is a trade-off between between spending time on irrelevant PAMs and possibly diffusing away, thereby missing the target.

This mechanism may also be applied to other protein that searches its target by a motif sequence. In practice, the real motif energy does not necessarily take the theoretical optimal value, due to biological or chemical constrains on the real value (for example, the interaction of Cas9 with the PAM is though hydrogen bonds, and hydrogen bonds has certain energy). However, a hypothesis can be made: the real motif energy is not likely to be very far away from the optimal value. Both the prediction of the dependence of the average search time on motif binding energy, and this hypothesis, may be tested by future experiments. In the case of Cas9, there is no experimental result yet about its average 3D diffusion time $\langle t_{3D} \rangle$ in a generic in vivo search sce-

nario, so we cannot compare our predicted PAM energy with the optimal energy. This comparison will be possible when measurements are made in the future.

Finally, the preliminary work in the last chapter studied the energetics and kinetics of the Cas9 recognition of the main target. The nucleic acid thermodynamic parameters predicted in section 5.1 may be compared with future experimental results in the same or similar ionic condition. These preliminary work may be continued, to build a more comprehensive model of the specificity of Cas9 than existing works. This will deepen our understanding on its behaviour when there are mismatches in the target, which is crucial for applications.

Bibliography

[Anders et al., 2014] Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of pam-dependent target dna recognition by the cas9 endonuclease. *Nature*, 513(7519):569–573.

[Anderson, 1958] Anderson, P. W. (1958). Absence of diffusion in certain random lattices. *Physical review*, 109(5):1492.

[Barrangou et al., 2007] Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007). Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712.

[Bauer et al., 2015] Bauer, M., Rasmussen, E. S., Lomholt, M. A., and Metzler, R. (2015). Real sequence effects on the search dynamics of transcription factors on dna. *Sci. Rep.*, 5(1):1–14.

[Berg et al., 1981] Berg, O. G., Winter, R. B., and Von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. models and theory. *Biochemistry*, 20(24):6929–6948.

[Biddle et al., 2011] Biddle, J., Priour Jr, D. J., Wang, B., and Sarma, S. D. (2011). Localization in one-dimensional lattices with non-nearest-neighbor hopping: Generalized anderson and aubry-andré models. *Phys. Rev. B*, 83(7):075105.

[Bonomo and Deem, 2018] Bonomo, M. E. and Deem, M. W. (2018). The physicist's guide to one of biotechnology's hottest new topics: Crispr-cas. *Physical biology*, 15(4):041002.

[Borland, 1963] Borland, R. (1963). The nature of the electronic states in disordered one-dimensional systems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 274(1359):529–545.

[Boyle et al., 2017] Boyle, E. A., Andreasson, J. O., Chircus, L. M., Sternberg, S. H., Wu, M. J., Guegler, C. K., Doudna, J. A., and Greenleaf, W. J. (2017). High-throughput biochemical profiling reveals sequence determinants of dcas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences*, 114(21):5461–5466.

[Brouns et al., 2008] Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V., and Van

Der Oost, J. (2008). Small crispr rnas guide antiviral defense in prokaryotes. *Science*, 321(5891):960–964.

[Cencini and Pigolotti, 2018] Cencini, M. and Pigolotti, S. (2018). Energetic funnel facilitates facilitated diffusion. *Nucleic acids research*, 46(2):558–567.

[Crisanti et al., 2012] Crisanti, A., Paladin, G., and Vulpiani, A. (2012). *Products of random matrices: in Statistical Physics*, volume 104. Springer Science & Business Media.

[Dahirel et al., 2009] Dahirel, V., Paillusson, F., Jardat, M., Barbi, M., and Victor, J.-M. (2009). Nonspecific dna-protein interaction: why proteins can diffuse along dna. *Physical review letters*, 102(22):228101.

[Dean and Bacon, 1963] Dean, P. and Bacon, M. (1963). The nature of vibrational modes in disordered systems. *Proceedings of the Physical Society*, 81(4):642.

[Dillard et al., 2018] Dillard, K. E., Brown, M. W., Johnson, N. V., Xiao, Y., Dolan, A., Hernandez, E., Dahlhauser, S. D., Kim, Y., Myler, L. R., Anslyn, E. V., et al. (2018). Assembly and translocation of a crispr-cas primed acquisition complex. *Cell*, 175(4):934–946.

[Eslami-Mossallam et al., 2022] Eslami-Mossallam, B., Klein, M., Smagt, C. V., Sanden, K. V., Jones Jr, S. K., Hawkins, J. A., Finkelstein, I. J., and Depken, M. (2022). A kinetic model predicts sp cas9 activity, improves off-target classification, and reveals the physical basis of targeting fidelity. *Nature communications*, 13(1):1367.

[Farasat and Salis, 2016] Farasat, I. and Salis, H. M. (2016). A biophysical model of crispr/cas9 activity for rational design of genome editing and gene regulation. *PLoS Comput. Biol.*, 12(1):e1004724.

[Feng et al., 2021] Feng, H., Guo, J., Wang, T., Zhang, C., and Xing, X.-h. (2021). Guide-target mismatch effects on dcas9–sgrna binding activity in living bacterial cells. *Nucleic Acids Research*, 49(3):1263–1277.

[Fineran and Charpentier, 2012] Fineran, P. C. and Charpentier, E. (2012). Memory of viral infections by crispr-cas adaptive immune systems: acquisition of new information. *Virology*, 434(2):202–209.

[Furstenberg, 1963] Furstenberg, H. (1963). Noncommuting random products. *Transactions of the American Mathematical Society*, 108(3):377–428.

[Furstenberg, 1971] Furstenberg, H. (1971). Random walks and discrete subgroups of lie groups. *Advances in probability and related topics*, 1:1–63.

[Gasiunas et al., 2012] Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9–crrna ribonucleoprotein complex mediates specific dna cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences*, 109(39):E2579–E2586.

[Globyte et al., 2019] Globyte, V., Lee, S. H., Bae, T., Kim, J.-S., and Joo, C. (2019). Crispr/cas9 searches for a protospacer adjacent motif by lateral diffusion. *The EMBO journal*, 38(4).

[Gong et al., 2018] Gong, S., Yu, H. H., Johnson, K. A., and Taylor, D. W. (2018). Dna unwinding is the primary determinant of crispr-cas9 activity. *Cell reports*, 22(2):359–371.

[Hachmo and Amir, 2022] Hachmo, O. and Amir, A. (2022). Conditional probability as found in nature: Facilitated diffusion. *arXiv preprint arXiv:2209.00500*.

[Hammar et al., 2012] Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E. G., Berg, O. G., and Elf, J. (2012). The lac repressor displays facilitated diffusion in living cells. *Science*, 336(6088):1595–1598.

[Herbert and Jones, 1971] Herbert, D. and Jones, R. (1971). Localized states in disordered systems. *Journal of Physics C: Solid State Physics*, 4(10):1145.

[Hille et al., 2018] Hille, F., Richter, H., Wong, S. P., Bratovič, M., Ressel, S., and Charpentier, E. (2018). The biology of crispr-cas: backward and forward. *Cell*, 172(6):1239–1259.

[Ishii, 1973] Ishii, K. (1973). Localization of eigenstates and transport phenomena in the one-dimensional disordered system. *Progress of Theoretical Physics Supplement*, 53:77–138.

[Ishino et al., 1987] Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in escherichia coli, and identification of the gene product. *Journal of bacteriology*, 169(12):5429–5433.

[Ivanov et al., 2020] Ivanov, I. E., Wright, A. V., Cofsky, J. C., Aris, K. D. P., Doudna, J. A., and Bryant, Z. (2020). Cas9 interrogates dna in discrete steps modulated by mismatches and supercoiling. *Proceedings of the National Academy of Sciences*, 117(11):5853–5860.

[Jansen et al., 2002] Jansen, R., van Embden, J. D., Gaastra, W., and Schouls, L. M. (2002). Identification of a novel family of sequence repeats among prokaryotes. *Omics: a journal of integrative biology*, 6(1):23–33.

[Jones et al., 2017] Jones, D. L., Leroy, P., Unoson, C., Fange, D., Ćurić, V., Lawson, M. J., and Elf, J. (2017). Kinetics of dcas9 target search in escherichia coli. *Science*, 357(6358):1420–1424.

[Josephs et al., 2015] Josephs, E. A., Kocak, D. D., Fitzgibbon, C. J., McMenemy, J., Gersbach, C. A., and Marszalek, P. E. (2015). Structure and specificity of the rna-guided endonuclease cas9 during dna interrogation, target binding and cleavage. *Nucleic acids research*, 43(18):8924–8941.

[Khakimzhan et al., 2020] Khakimzhan, A., Garenne, D., and Tickman, B. I. (2020). Proofreading mechanism of class 2 crisprcas systems. *arXiv*, 20(24):6929–6948.

[Klein et al., 2018] Klein, M., Eslami-Mossallam, B., Arroyo, D. G., and Depken, M. (2018). Hybridization kinetics explains crispr-cas off-targeting rules. *Cell reports*, 22(6):1413–1423.

[Lomholt et al., 2009] Lomholt, M. A., van den Broek, B., Kalisch, S.-M. J., Wuite, G. J., and Metzler, R. (2009). Facilitated diffusion with dna coiling. *Proc. Natl. Acad. Sci.*, 106(20):8204–8208.

[Lu et al., 2021] Lu, Q., Bhat, D., Stepanenko, D., Pigolotti, S., et al. (2021). Search and localization dynamics of the crispr-cas9 system. *Physical Review Letters*, 127(20):208102.

[MacQuarrie et al., 2011] MacQuarrie, K. L., Fong, A. P., Morse, R. H., and Tapscott, S. J. (2011). Genome-wide transcription factor binding: beyond direct target regulation. *Trends in Genetics*, 27(4):141–148.

[Makarova et al., 2017a] Makarova, K. S., Zhang, F., and Koonin, E. V. (2017a). Snapshot: class 1 crispr-cas systems. *Cell*, 168(5):946–946.

[Makarova et al., 2017b] Makarova, K. S., Zhang, F., and Koonin, E. V. (2017b). Snapshot: class 2 crispr-cas systems. *Cell*, 168(1):328–328.

[Marklund et al., 2022] Marklund, E., Mao, G., Yuan, J., Zikrin, S., Abdurakhmanov, E., Deindl, S., and Elf, J. (2022). Sequence specificity in dna binding is mainly determined by association rather than dissociation. *Biophysical Journal*, 121(3):480a.

[Martens et al., 2019] Martens, K. J., van Beljouw, S. P., van der Els, S., Vink, J. N., Baas, S., Vogelaar, G. A., Brouns, S. J., van Baarlen, P., Kleerebezem, M., and Hohlbein, J. (2019). Visualisation of dcas9 target search in vivo using an open-microscopy framework. *Nature communications*, 10(1):1–11.

[Matsuda and Ishii, 1970] Matsuda, H. and Ishii, K. (1970). Localization of normal modes and energy transport in the disordered harmonic chain. *Progress of Theoretical Physics Supplement*, 45:56–86.

[Mirny et al., 2009] Mirny, L., Slutsky, M., Wunderlich, Z., Tafvizi, A., Leith, J., and Kosmrlj, A. (2009). How a protein searches for its site on dna: the mechanism of facilitated diffusion. *Journal of Physics A: Mathematical and Theoretical*, 42(43):434013.

[Mojica et al., 2009] Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic crispr defence system. *Microbiology*, 155(3):733–740.

[Mojica et al., 2000] Mojica, F. J., Díez-Villaseñor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of archaea, bacteria and mitochondria. *Molecular microbiology*, 36(1):244–246.

[Redner, 2002] Redner, S. (2002). A guide to first-passage processes.

[Rhee and Pugh, 2011] Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419.

[Riggs et al., 1970] Riggs, A. D., Bourgeois, S., and Cohn, M. (1970). The lac represser-operator interaction: Iii. kinetic studies. *Journal of molecular biology*, 53(3):401–417.

[SantaLucia Jr, 1998] SantaLucia Jr, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4):1460–1465.

[Sheinman et al., 2012] Sheinman, M., Bénichou, O., Kafri, Y., and Voituriez, R. (2012). Classes of fast and specific search mechanisms for proteins on dna. *Reports on Progress in Physics*, 75(2):026601.

[Singh et al., 2016] Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A., and Ha, T. (2016). Real-time observation of dna recognition and rejection by the rna-guided endonuclease cas9. *Nature communications*, 7(1):1–8.

[Singh et al., 2018] Singh, D., Wang, Y., Mallon, J., Yang, O., Fei, J., Poddar, A., Ceylan, D., Bailey, S., and Ha, T. (2018). Mechanisms of improved specificity of engineered cas9s revealed by single-molecule fret analysis. *Nature structural & molecular biology*, 25(4):347–354.

[Slutsky and Mirny, 2004] Slutsky, M. and Mirny, L. A. (2004). Kinetics of protein-dna interaction: facilitated target location in sequence-dependent potential. *Biophysical journal*, 87(6):4021–4035.

[Sternberg et al., 2014] Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., and Doudna, J. A. (2014). Dna interrogation by the crispr rna-guided endonuclease cas9. *Nature*, 507(7490):62–67.

[Szczelkun et al., 2014] Szczelkun, M. D., Tikhomirova, M. S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of r-loop formation by single rna-guided cas9 and cascade effector complexes. *Proceedings of the National Academy of Sciences*, 111(27):9798–9803.

[Thouless, 1972] Thouless, D. (1972). A relation between the density of states and range of localization for one dimensional random systems. *Journal of Physics C: Solid State Physics*, 5(1):77.

[Van Der Oost et al., 2014] Van Der Oost, J., Westra, E. R., Jackson, R. N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of crispr–cas systems. *Nature Reviews Microbiology*, 12(7):479–492.

[Westra et al., 2012] Westra, E. R., Swarts, D. C., Staals, R. H., Jore, M. M., Brouns, S. J., and van der Oost, J. (2012). The crisprs, they are a-changin': how prokaryotes generate adaptive immunity. *Annual review of genetics*, 46:311–339.

[Winter et al., 1981] Winter, R. B., Berg, O. G., and Von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. the escherichia coli lac repressor-operator interaction: kinetic measurements and conclusions. *Biochemistry*, 20(24):6961–6977.

[Winter and Von Hippel, 1981] Winter, R. B. and Von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. the escherichia coli lac repressor-operator interaction: equilibrium measurements. *Biochemistry*, 20(24):6948–6960.

[Zwanzig, 1988] Zwanzig, R. (1988). Diffusion in a rough potential. *Proceedings of the National Academy of Sciences*, 85(7):2029–2030.

Appendix A

Maximum likelihood fit

This appendix is the same as part 1 of the supplemental material of [Lu et al., 2021]. To fit the experimental data from [Globyte et al., 2019], we express the likelihood of one specific experiment as

$$\mathcal{L}_j = \mathcal{N}_j! \prod_i \frac{\rho_{i,j}^{n_{i,j}}}{n_{i,j}!} \tag{A.1}$$

where the index $0 \leq j \leq 5$ indicates the number of PAM sites in each experiment. For each experiment j, we call $n_{i,j}$ the number of binding events in the ith bin of the histogram, $\mathcal{N}_j = \sum_i n_{i,j}$ is the total number of binding events, and $\rho_{i,j} = P(t_{i-1}) - P(t_i)$ is the probability that the duration of a binding event falls into the ith bin. This probability is obtained from numerical integration of Eq. (1) in the Main Text for a given choice of the parameters k, D, and β, with a matrix $\hat{A}$ determined by the arrangement of PAM sites in the given experiment. We maximize the joint log-likelihood

$$\ln \mathcal{L} = \sum_{j=0}^{5} \ln \mathcal{L}_j \tag{A.2}$$

with respect to the three parameters and compute their uncertainties from the curvature of the log-likelihood. To facilitate a visual comparison, individual curves for each number of PAM sites and corresponding experimental data are shown in Fig. A.1 (same as Fig. 2b in the Main Text, but with each experiment in a different panel).

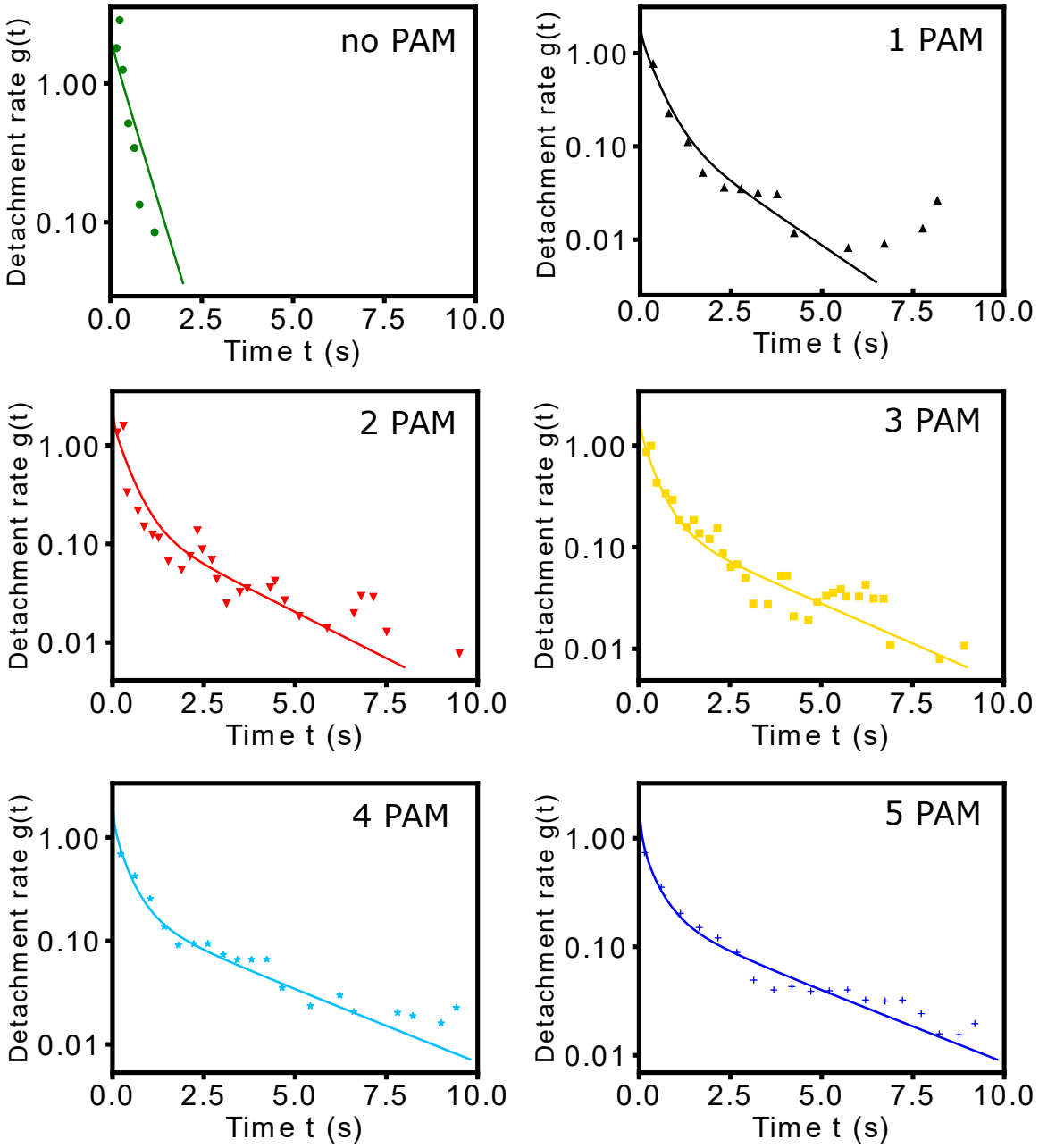


Figure A.1: Fitted detachment rate as a function of time for different number of PAMs. Curves and data are the same as in Fig. 2b of the Main Text, but presented in separate panels.

Appendix B

Sequence-dependent model

This appendix is the same as part 2 of the supplemental material of [Lu et al., 2021]. In the model introduced in the Main Text, the binding energy of Cas9 to any triplet other than PAM is the same. In this Section, we introduce a model that relaxes this assumption and study its properties. To this aim, we define as "canonical PAM" a NGG triplet (where N stands for any base) and "non-canonical PAMs" the 15 possible triplets where either one or both G are replaced by other bases. This definition is motivated by the observation that the first "N" base of PAM does not to affect the binding energy of Cas9 [Bonomo and Deem, 2018]. However, in principle, the binding energy of Cas9 with each non-canonical PAM can depend on the other two bases, and the assumption made in the Main Text should be considered as a simplification.

We determine the binding energies of non-canonical PAMs from experimental results of the equilibrium occupancy of off-target dsDNA bound by dCas9 [Boyle et al., 2017], a mutant of Cas9 that lacks the endonuclease capability. In the experiment, double strand DNA sequences containing the 20 bps main target and all possible replacement of the "GG" in the PAM are fixed in the flow cell. After 12 hours incubation using 10nM dCas9, the occupancy of the dsDNA sequences is measured (see Fig. 2S in [Boyle et al., 2017]).

The measurement is performed after incubation, so that the system can be assumed to be at equilibrium. Every target DNA sequence can either be occupied by one Cas9 or empty. The occupancy O is therefore expressed by the Fermi-Dirac distribution

$$O_i = \frac{1}{1 + e^{\epsilon_i - \mu}} \tag{B.1}$$

where ϵ_i is the binding energy of a particular triplet i and μ is the chemical potential of dCas9. In the experiment, all canonical and non-canonical PAMs are followed by an identical 20bp target. Therefore, we expect differences in ϵ_i to depend on the different non-canonical PAMs only. The authors of Ref. [Boyle et al., 2017] report the occupancy O_i for all non-canonical PAMs relative to the canonical one. We call ϵ_T the binding energy of the specific target (i.e. the canonical PAM). We assume that ϵ_T is sufficiently negative so that the occupancy of the target is approximately equal to 1. Accordingly, we interpret the relative occupancies O_i as absolute ones.

We use Eq. (B.1) to eliminate μ and express the binding energy difference between a generic non-canonical PAM and the weakest non-canonical PAM (NTC) that we take

as reference:

$$\Delta\epsilon_i = \epsilon_i - \epsilon_{NTC} = \ln\left[\frac{O_{NTC}(1-O_i)}{O_i(1-O_{NTC})}\right]. \tag{B.2}$$

Equation (B.2) permits to determine the binding energy difference from the experimental occupancy data. Results are presented in Table B.1.

The diffusion and unbinding rates of the sequence-dependent model are defined from these energy differences as:

$$\begin{aligned} D_{n+1,n} &= D_{n-1,n} = D' e^{\Delta\epsilon_n} \\ k_n &= k' e^{\Delta\epsilon_n}, \end{aligned} \tag{B.3}$$

where we denoted the diffusion rate and the unbinding rate with D' and k', respectively, to distinguish them from the rate D and k appearing in the model presented in the Main Text. We run simulations of the sequence-dependent model using the binding energy differences in Table B.1. In this case, the free parameters are D', k', and the binding energy difference $\Delta\epsilon_T = \epsilon_T - \epsilon_{NTC}$ between the canonical PAM and the weakest non-canonical NTC. We fit these three parameters using the FRET data from [Boyle et al., 2017], following the same procedure described in Section I and using the specific DNA sequences that Ref. [Boyle et al., 2017] reports for each experiment. We obtain $D' = 160s^{-1}, k' = 6.57s^{-1}$, and $\Delta\epsilon_T = -4.47$.

	G	A	C	T
G		-2.61	-1.12	-1.42
A	-2.59	-1.04	-0.975	-1.35
C	-1.22	-1.40	-1.35	0
T	-1.08	-0.953	-0.680	-1.12

Table B.1: non-canonical PAM energies $\Delta\epsilon_i$. Rows represent the first nucleotide and columns for the nucleotide next to the "N"

We now compare these parameters with those for the nearest-neighbor model presented in the Main Text. In the sequence-dependent model, the average energy of non-canonical PAM sites is $\epsilon_{av} = -1.26$ (see Table I). We now express the average diffusion rate between neighboring non-canonical PAM sites as

$$\langle D_{n+1,n}\rangle = D'\langle e^{\Delta\epsilon_n}\rangle = 54s^{-1} \tag{B.4}$$

In contrast, in the model presented in the Main Text, we have $\langle D_{n+1,n}\rangle = D = 52s^{-1}$. The relative difference between these two values is about 4%.

In the sequence-dependent model, we similarly have that the average unbinding rate from a non-canonical PAM is expressed by

$$\langle k_n\rangle = k'\langle e^{\Delta\epsilon_n}\rangle = 2.2s^{-1} \tag{B.5}$$

whereas in the model of the Main Text we have $\langle k_n\rangle = k = 1.94s^{-1}$. In this case, the relative discrepancy is 12.5%.

Finally, in the sequence-dependent model, the energy difference between the canon-

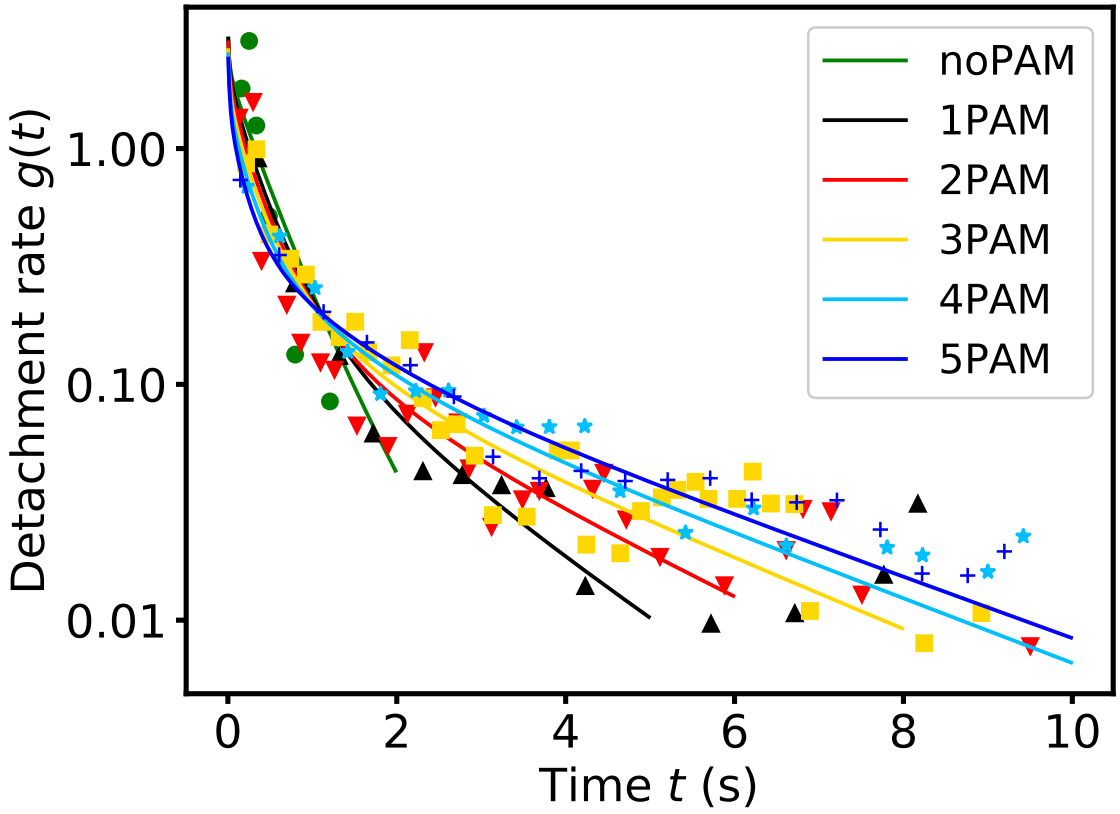


Figure B.1: Detachment rate $g(t)$ for $j = 0 \ldots 5$ PAM sites predicted by the sequence-dependent model (lines) versus experimental measures from Ref. [Boyle et al., 2017] (points). See Fig. 2b in the Main Text for comparison and more information. The fit returns a value of $\chi^2 = 280.4$, compared with $\chi^2 = 276.6$ in the model presented in the Main Text.

ical PAM and an average non-canonical PAM is equal to $\epsilon_T - \epsilon_{av} = -3.21$. This value is close to the estimated value $\beta = -3.34$ of the model in the Main Text, with a relative discrepancy of 4%.

With these fitted parameters, we find that the sliding length in the one PAM case is equal to $\ell = 5.2$ bp compared with 6.2 bp for the model in the Main Text.

We also computed the localization length and the density of states for the sequence-dependent model in the disordered case, see Fig. B.2. For the sequence-dependent model, the maximum localization length is slightly larger than for the model in the Main Text ($\gamma \approx 15$ vs $\gamma \approx 10$, respectively). This difference should not be surprising, since the localization length is expected to be particularly sensitive to the distribution of the disorder. In any case, the qualitative result is confirmed, in the sense that both sliding lengths are much shorter than what it is observed in immunoprecipitation experiments.

We conclude from these comparisons that the physical picture resulting from the model presented in the Main Text is consistent with the one provided by this more detailed model.

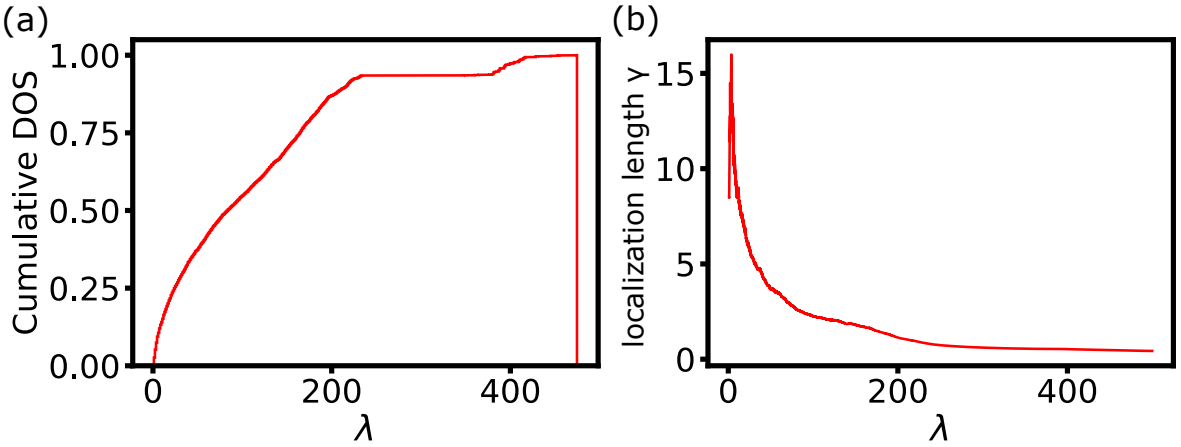


Figure B.2: (a) Cumulative density of states (DOS) and (b) localization length as function of λ for the sequence-dependent model, Eq. (B.3), computed the transfer matrix method and Eqs. (3.14) and (3.15) in the Main Text. The DNA chain length is $N = 5000$.

Appendix C

Regular versus disordered assortment of PAM Sites

This appendix is the same as part 3 of the supplemental material of [Lu et al., 2021]. Our interpretation of facilitated diffusion of Cas9 as a localization phenomenon leads to an interesting prediction. We expect eigenvectors characterizing Cas9 dynamics on a long DNA chain to be localized only if the PAM sites are arranged in a disordered fashion. If, instead, the PAM sites are regularly spaced, the eigenvectors should be extended as there is no disorder in this case. This prediction is confirmed in Fig. C.1. The figure shows that, in the case of regularly spaced PAM sites, the eigenvectors are characterized by peaks at each PAM site modulated by wave-like envelopes spanning the entire system size.

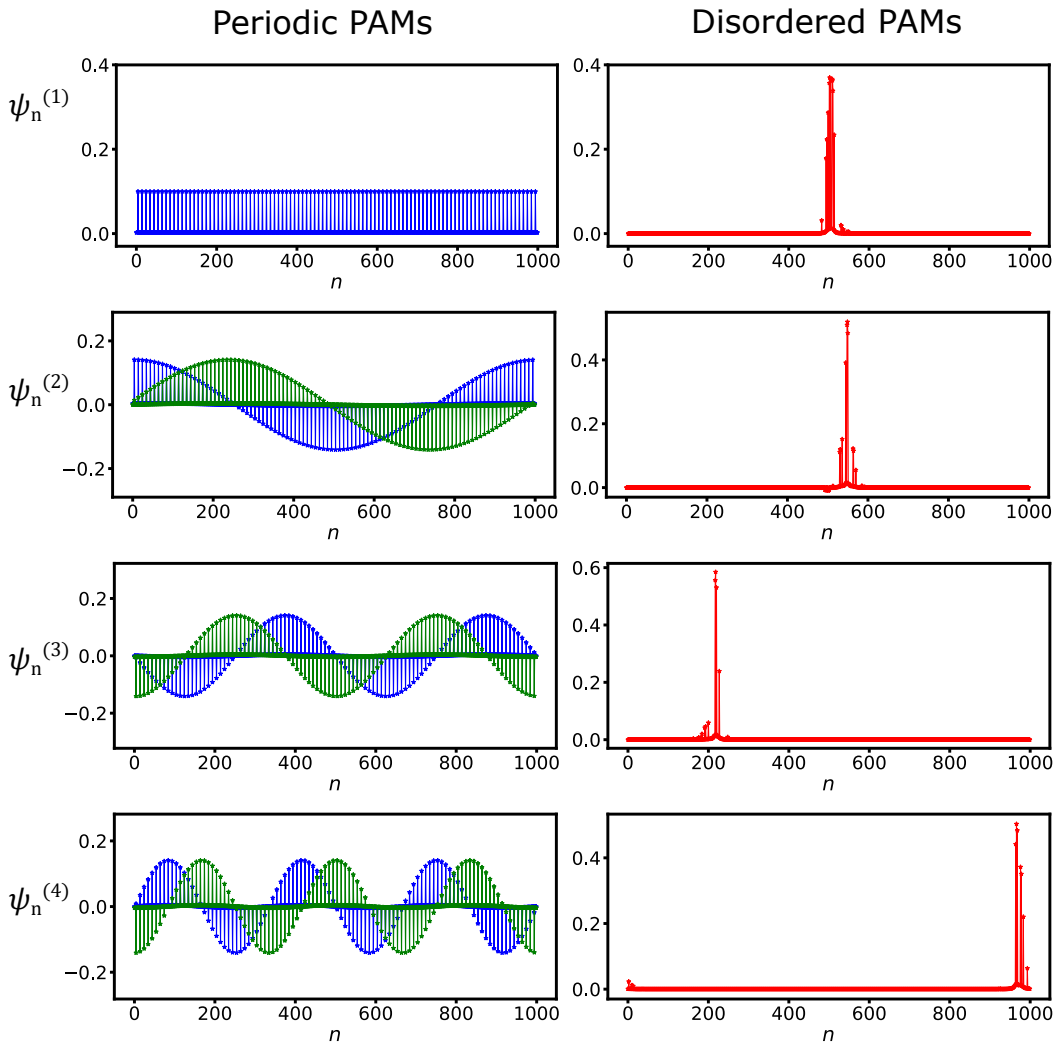


Figure C.1: Comparison of the first four eigenvectors of Cas9 sliding dynamics for (left) periodically spaced PAMs and (right) a disordered arrangement of PAM sites. In both cases, the length of the DNA chain is $N = 1000$ and the average density of PAM sites is $1/10$. In the periodic case, the eigenvalues λ_2, λ_3, and λ_4 are associated with two degenerate eigenvectors (shown in blue and green in the figures). We obtained qualitatively similar results for closed boundary conditions (not shown).

Appendix D

Hopping model

This appendix is the same as part 4 of the supplemental material of [Lu et al., 2021].

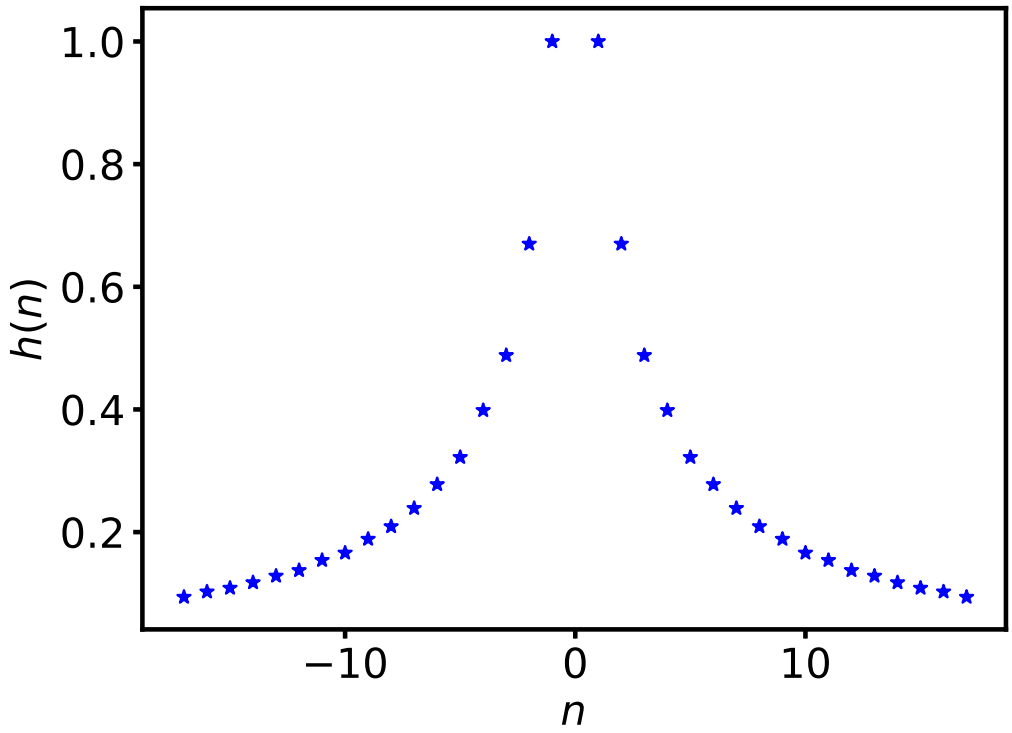


Figure D.1: Plot of the hopping distribution $h(n)$ versus n for $\alpha = 1$. The distribution $h(n)$ is normalized so that $h(1) = 1$ and truncated at $n = 17$ for computational convenience.

The hopping distribution $h(n)$ can be estimated from the solution of a diffusion equation in cylindrical coordinates [Lomholt et al., 2009]. The assumption of cylindrical symmetry is justified as far as we limit ourselves to hopping at distances much shorter than the DNA persistence length, which is on the order of 150 base pairs. On these short distances, the DNA double helix can be regarded as a straight cylinder.

We consider the probability $W(n, t)$ of a protein to rebind at coordinate n at time

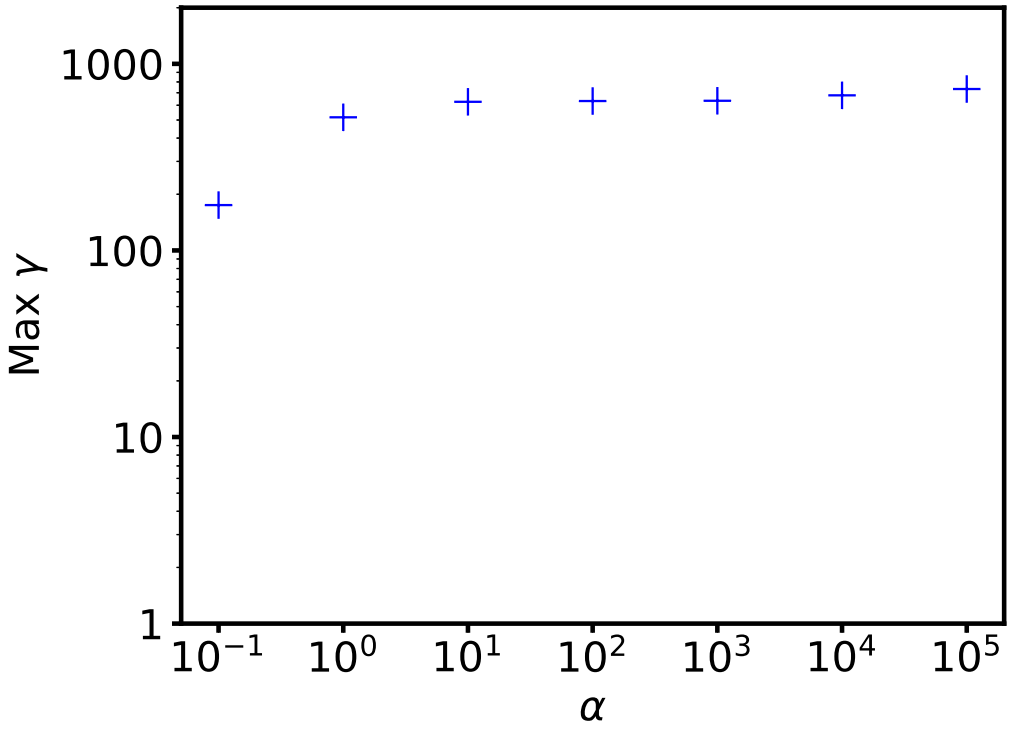


Figure D.2: Maximum localization length in the spectrum as a function of α. For each value of α, the maximum localization length is computed by direct diagonalization (as in Fig. 4d of the Main Text).

t, given that it detached at position $n = 0$ and $t = 0$. From the diffusion equation in cylindrical coordinates, the authors of Ref. [Lomholt et al., 2009] obtains the Fourier-Laplace transform

$$\tilde{W}(q,u) = \int_0^\infty dt\ e^{-ut} \int_{-\infty}^{\infty} dn\ e^{iqn} W(n,t). \tag{D.1}$$

In particular, the Fourier-Laplace transform calculated in $u = 0$ yields the Fourier transform of the integrated probability of hopping to a given distance at any time:

$$\begin{aligned} \tilde{W}(q,0) &= \int_0^\infty dt \int_{-\infty}^{\infty} dn\ e^{iqn} W(n,t) \\ &= \left[1 + \frac{2\pi\alpha|q|rK_1(|q|r)}{(K_0(|q|r))}\right]^{-1}, \end{aligned} \tag{D.2}$$

where $r = 3$ is the DNA radius, measured in unit of the base pair distance, $K_j(n)$ is the modified Bessel function of the second kind, and $\alpha = 1$ is the ratio between the 3D diffusion coefficient and the non-specific binding rate. For $n > 1$, we compute $h(n)$ by a numerical inverse Fourier transform of $\tilde{W}(q,0)$, truncated at a distance $n_{\max} = 17$. This maximum distance is chosen for computational convenient and is consistent with the assumption of cylindrical symmetry, as previously discussed. The hopping distribution

$h(n)$ for $\alpha = 1$, normalized such that $h(1) = 1$, is shown in Fig. D.1. The localization length for the hopping model for $\alpha = 1$ is shown in Fig. 4 of the Main Text. We found qualitatively similar results for the sliding length for α ranging from 0.1 to 10^5, see Fig. D.2.

Appendix E

Derivation of Equation (4.20) and (4.21)

E.0.1 The first type of trajectories

In the first type of trajectories, between the first binding to the central PAM and the final recognition, there is no detachment and 3D diffusion. The trajectories in this type can be divided further into different groups by their number of times of visiting the PAM at $x = 0$, from one to infinity.

In the first group the $x = 0$ PAM is visited only once, so the Cas9 transits into the recognition mode directly. A particular event in this group would be: the Cas9 reaches the central PAM in the time interval T_0 to $T_0 + dT_0$, then transits to the R-mode in time $T_0 + \tau$ to $T_0 + \tau + d\tau$. The probability of the former is $P_0(T_0)dT_0$, and that of the latter is $\frac{p}{f^T+2De^\beta}e^{-(f^T+2De^\beta)\tau}d\tau$. The total time spent is $T_0 + \tau$. Since these two processes are independent from each other, the combined probability is the product of the two probabilities. Therefore, the contribution of this particular event to $\langle T_{tot}\rangle$ is $(T_0 + \tau)P_0(T_0)dT_0\frac{p}{f^T+2De^\beta}e^{-(f^T+2De^\beta)\tau}d\tau$. And the contribution of the first group in sum to $\langle T_{tot}\rangle$ is given by the integral

$$\int_0^{+\infty}\int_0^{+\infty}(T_0+\tau)P_0(T_0)dT_0\frac{p}{f^T+2De^\beta}e^{-(f^T+2De^\beta)\tau}d\tau = p\langle T_0\rangle + \frac{p}{f^T+2De^\beta}. \quad \text{(E.1)}$$

The second group of the first type corresponds to those trajectories that visited the central PAM twice. A particular event in this group would be: the Cas9 reaches the central PAM in the time interval T_0 to $T_0 + dT_0$, then diffuses to one of its neighbours within $T_0 + \tau_1'$ to $T_0 + \tau_1' + d\tau_1'$, then returns to the central PAM within $T_0 + \tau_1' + t$ to $T_0+\tau_1'+t+dt$, finally transits to the R-mode within $T_0+\tau_1'+t+\tau$ to $T_0+\tau_1'+t+\tau+d\tau$. The probabilities for these four processes are $P_0(T_0)dT_0$, $\frac{1-p}{f^T+2De^\beta}e^{-(f^T+2De^\beta)\tau_1'}d\tau_1'$, $g(t)dt$ and $\frac{p}{f^T+2De^\beta}e^{-(f^T+2De^\beta)\tau}d\tau$, respectively, in which $g(t)$ is the first passage time distribution from $n = \pm 1$ to $n = 0$ taken the possible detachment also into account. In order to calculate the contribution from this group, we first need to solve $g(t)$.

A standard result in continuous space random walk is that the first passage time distribution from n_0 to the origin is $\frac{n_0}{\sqrt{4\pi Dt^3}}e^{-\frac{n_0^2}{4Dt}}$. The counterpart in discrete space random walk (as in our model) is in terms of the modified Bessel function of the first kind [Redner, 2002]. Therefore we make the approximation by using the result in

continuous space RW here, to simplify our calculation. A simulations was done to check this approximation, with analytical prediction from continuous space RW compared with numerical results from discrete space RW. The result shows the approximation works well. The only difference from the a net first passage time problem is that, in our case we require that it returns to $n = 0$ before detaching, so $g(t)$ is obtained by letting $n_0 = 1$ in the previous equation and decreasing it by a factor of e^{-kt} (the probability that it is still on the DNA). Therefore $g(t)$ is not normalized to 1:

$$g(t) = \frac{1}{\sqrt{4\pi D t^3}} e^{-\frac{1}{4Dt} - kt}. \tag{E.2}$$

Here, there is a assumption that in this 1D diffusion process from $n = \pm 1$ to the first return at $n = 0$, the diffusion and detachment rates are always D and k, but I will show in the end that this can be relaxed. We have

$$\int_0^{+\infty} g(t)dt = \frac{1}{\sqrt{4\pi D t^3}} e^{-\frac{1}{4Dt} - kt} dt = e^{-\sqrt{\frac{k}{D}}}, \tag{E.3}$$

and

$$\langle t \rangle = \int_0^{+\infty} t g(t)dt = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{1}{4Dt} - kt} dt = \frac{e^{-\sqrt{\frac{k}{D}}}}{\sqrt{4Dk}}. \tag{E.4}$$

Now we calculate the contribution of the second group to $\langle T_{tot} \rangle$:

$$\begin{aligned} &\int_0^{+\infty}\int_0^{+\infty}\int_0^{+\infty}\int_0^{+\infty} (T_0 + \tau_1' + t + \tau) P_0(T_0) dT_0 \\ &\frac{1-p}{f^T + 2De^\beta} e^{-(f^T + 2De^\beta)\tau_1'} d\tau_1' g(t) dt \frac{p}{f^T + 2De^\beta} e^{-(f^T + 2De^\beta)\tau} d\tau \\ &= p\langle T_0 \rangle (1-p) e^{-\sqrt{\frac{k}{D}}} + \frac{p}{f^T + 2De^\beta}(1-p) e^{-\sqrt{\frac{k}{D}}} \\ &+ p\frac{e^{-\sqrt{\frac{k}{D}}}}{\sqrt{4Dk}}(1-p) + \frac{p}{f^T + 2De^\beta}(1-p) e^{-\sqrt{\frac{k}{D}}} \end{aligned} \tag{E.5}$$

The terms in the result are in the order of the corresponding processes.

In the third group, the Cas9 visits the central PAM three times before the R-mode. Comparing with the second group, there is one more process of "staying at the central PAM and diffuse away to $n = \pm 1$" that costs τ_2', and one more process of returning to $n = 0$ from $n = \pm 1$. The total time is $(T_0 + \tau_1' + t_1 + \tau_2' + t_2 + \tau)$, the calculation is similar, but with 6 integrations. The result is

$$\begin{aligned} &p\langle T_0\rangle \left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)^2 + 2\frac{p}{f^T+2De^\beta}\left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)^2 \\ &+2p\frac{1}{\sqrt{4Dk}}\left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)^2 + \frac{p}{f^T+2De^\beta}\left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)^2 \end{aligned} \tag{E.6}$$

The rest groups are calculated in the similar way, the number of terms in the final results stays 4, since the added two new processes' duration always join the middle two terms. Now by inspection of Eq. E.1, E.5, E.6, we know that there are 4 series in the contribution of the first type trajectories: the first and last terms in these equations correspond to two geometric series Σr^n, and the middle terms in Eq. E.5, E.6 correspond to two series of the type $\Sigma n r^n$, where $r = \left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)$. Therefore, the contribution from the first type as a whole is

$$p\langle T_0\rangle\frac{1}{1-r} + \frac{p(1-p)e^{-\sqrt{\frac{k}{D}}}}{f^T+2De^\beta}\frac{1}{(1-r)^2} + \frac{p(1-p)e^{-\sqrt{\frac{k}{D}}}}{\sqrt{4Dk}}\frac{1}{(1-r)^2} + \frac{p}{f^T+2De^\beta}\frac{1}{1-r} \tag{E.7}$$

E.0.2 The second type of trajectories

In this type, Cas9 detaches at least once between its first encounter with the central PAM and its final arrival to the R-mode. The trajectories in this type can also be divided further into different groups by their number of times of visiting the PAM at $x = 0$ before detach, from just one to infinity.

A particular event in the first group would be: the Cas9 reaches the central PAM in the time interval T_0 to $T_0 + dT_0$, then diffuses to one of its neighbours within $T_0 + \tau_1'$ to $T_0 + \tau_1' + d\tau_1'$, then detaches within $T_0 + \tau_1' + t'$ to $T_0 + \tau_1' + t' + dt'$ before returning to $n = 0$, then makes a 3D diffusion of length t_{3D}, and finally spends another time T to reach the R-mode. The probabilities for these four processes are $P_0(T_0)dT_0$, $\frac{1-p}{f^T+2De^\beta}e^{-(f^T+2De^\beta)\tau_1'}d\tau_1'$, $h(t')dt'$, 1 (since t_{3D} is a constant), and $P(T)dT$, respectively. Here, $h(t')$ is the first passage time distribution from $n = \pm 1$ to detachment but before it returned to $n = 0$. In order to calculate the contribution from this group, we first need to solve $h(t')$. $P(T)$ is exactly the same distribution as $P(T_{tot})$, this is because once the Cas9 detaches, the whole search process starts again. The total time is $T_0 + \tau_1' + t' + t_{3D} + T$.

To calculate $h(t')$, first note that the detachment distribution is just $ke^{-kt'}$. But we also require it never returned to $n = 0$ during this process, so the distribution has to be decreased by a certain factor $G(t')$. To calculate $G(t')$, as in the calculation of $g(t)$, we use results from continuous space RW as a convenient approximation to avoid Bessel functions. The standard result is that the probability of the walker stays to the right of $x = 0$, given it started at $x = n_0$, is $G(t'; n_0) = 1 - \int_0^{t'} \frac{n_0}{\sqrt{4\pi D t^3}} e^{-\frac{n_0^2}{4Dt}} dt$. Using it in our case, we let $n_0 = 1$ as before, it reduces to $G(t') = erf\left(\frac{1}{\sqrt{4Dt'}}\right)$. So we have

$$h(t') = ke^{-kt'} erf\left(\frac{1}{\sqrt{4Dt'}}\right). \tag{E.8}$$

As in the last subsection, there is a assumption that in this 1D diffusion process from $n = \pm 1$ to detachment, the diffusion and detachment rates are always D and k, but again I will show in the end that this can be relaxed. We then have

$$\int_0^{+\infty} h(t')dt' = 1 - e^{-\sqrt{\frac{k}{D}}}. \tag{E.9}$$

Note that $\int_0^{+\infty} h(t')dt' + \int_0^{+\infty} g(t)dt = 1$, this is expected because essentially the former is the probability of detaching before reaching $n = 0$ and the latter is the probability of reaching $n = 0$ before detaching, any trajectory starts at $n = \pm 1$ must end up with those two.

Also

$$\langle t' \rangle = \int_0^{+\infty} t'h(t')dt' = \frac{1}{k}\left(1 - e^{-\sqrt{\frac{k}{D}}}\left(1 + \sqrt{\frac{k}{4D}}\right)\right). \tag{E.10}$$

The contribution of the first group is calculated by a quadruple integration as in the last subsection. The result is

$$\begin{aligned} &(\langle T_0 \rangle + t_{3D} + \langle T_{tot} \rangle)(1 - e^{-\sqrt{\frac{k}{D}}})(1-p) \\ +\frac{(1-p)}{f^T + 2De^{\beta}}(1 - e^{-\sqrt{\frac{k}{D}}}) &+ \frac{1}{k}\left(1 - e^{-\sqrt{\frac{k}{D}}}\left(1 + \sqrt{\frac{k}{4D}}\right)\right)(1-p). \end{aligned} \tag{E.11}$$

The $\langle T_{tot} \rangle$ is exactly what we want, and comes from the average of T ($\int_0^{+\infty} P(T)TdT = \int_0^{+\infty} T_{tot}P(T_{tot})dT_{tot} = \langle T_{tot} \rangle$) in the process. In the end $\langle T_{tot} \rangle$ will be solved as a unknown, see the main text. From now on the terms are not written in order of their happening since we can obviously combine terms with common factors.

A particular event in the second group visit $n = 0$ two times before detach. This adds one copy of the time spent returning to $n = 0$, i.e. t as denoted in the last subsection, and another copy of the time spent by diffusing away from $n = 0$, i.e. a τ_2' besides the τ_1' in the first group. The result is

$$
\begin{aligned}
&(\langle T_0\rangle + t_{3D} + \langle T_{tot}\rangle)(1-e^{-\sqrt{\frac{k}{D}}})(1-p)^2 e^{-\sqrt{\frac{k}{D}}} \\
&+2\frac{(1-p)}{f^T+2De^{\beta}}(1-e^{-\sqrt{\frac{k}{D}}})(1-p)e^{-\sqrt{\frac{k}{D}}} \\
+\frac{1-p}{\sqrt{4Dk}}(1-e^{-\sqrt{\frac{k}{D}}})(1-p)e^{-\sqrt{\frac{k}{D}}} &+ \frac{1}{k}\left(1-e^{-\sqrt{\frac{k}{D}}}\left(1+\sqrt{\frac{k}{4D}}\right)\right)(1-p)^2 e^{-\sqrt{\frac{k}{D}}}.
\end{aligned}
\tag{E.12}
$$

The calculation is similar for the third group, with the result

$$
\begin{aligned}
&(\langle T_0\rangle + t_{3D} + \langle T_{tot}\rangle)(1-e^{-\sqrt{\frac{k}{D}}})(1-p)\left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)^2 \\
&+3\frac{(1-p)}{f^T+2De^{\beta}}(1-e^{-\sqrt{\frac{k}{D}}})\left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)^2 \\
&+2\frac{1-p}{\sqrt{4Dk}}(1-e^{-\sqrt{\frac{k}{D}}})\left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)^2 \\
+\frac{1}{k}\left(1-e^{-\sqrt{\frac{k}{D}}}\left(1+\sqrt{\frac{k}{4D}}\right)\right)&(1-p)\left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)^2.
\end{aligned}
\tag{E.13}
$$

Similar to the first type, it is clear that there are also 4 infinite series, the first and last terms in Eq. E.11, E.12, E.13 correspond to two geometric series Σr^n, and the middle terms in Eq. E.12, E.13 correspond to two series of the type $\Sigma n r^n$, where $r = \left((1-p)e^{-\sqrt{\frac{k}{D}}}\right)$. Therefore, the contribution from the second type as a whole is

$$
\begin{aligned}
&(\langle T_0\rangle + t_{3D} + \langle T_{tot}\rangle)(1-e^{-\sqrt{\frac{k}{D}}})(1-p)\frac{1}{1-r} \\
&+\frac{(1-p)}{f^T+2De^{\beta}}(1-e^{-\sqrt{\frac{k}{D}}})\frac{1}{(1-r)^2} + \frac{(1-p)^2 e^{-\sqrt{\frac{k}{D}}}}{\sqrt{4Dk}}(1-e^{-\sqrt{\frac{k}{D}}})\frac{1}{(1-r)^2} \\
&+\frac{1}{k}\left(1-e^{-\sqrt{\frac{k}{D}}}\left(1+\sqrt{\frac{k}{4D}}\right)\right)(1-p)\frac{1}{1-r}
\end{aligned}
\tag{E.14}
$$