Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Faculty of Applied Sciences

Cleavage and unbinding times of CRISPR associated proteins

Thesis to obtain the degrees of

BACHELOR OF SCIENCE
in
APPLIED MATHEMATICS
and
APPLIED PHYSICS

by

LUUK VAN DUUREN

Delft, The Netherlands
July 2018

BSc thesis APPLIED MATHEMATICS and APPLIED PHYSICS

'Cleavage and unbinding times of CRISPR associated proteins'

LUUK VAN DUUREN

Delft University of Technology

Supervisors

Prof. dr. F.H.J. Redig

Dr. S.M. Depken
M. Klein

Other committee members

Dr. ir. W.G.M. Groenevelt

Dr. T. Idema

July, 2018

Delft

Abstract

In this thesis the CRISPR-Cas9 mechanism, a promising mechanism for gene-editing, is considered. Closed form expressions are derived for the probability and time to cleave or unbind for the associated Cas9 protein. The mechanism can be modelled mathematically by a birth and death process, therefore the expressions could be derived using Markov chains and semigroups. The expressions are compared to simulations and interpreted using the model of hybridization kinetics. Finally the moment generating function of the stopping time is derived for two special Markov processes, i.e. a random walk and a Brownian motion with drift. This was done using martingales.

Contents

1 Introduction

Organisms are regularly intruded by phages and viruses. Such intrusion could have negative effects on the hosting organism. Therefore many of them have found a way to protect themselves from these intruders. In many prokaryotes[1] this protection is organised by the CRISPR[2] immunity mechanism and its Cas9 (CRISPR-associated) proteins. This system integrates parts of the genome of intruding phages in the bacterial DNA, creating a record of infection [6]. Furthermore it creates Cas9 proteins which contain a piece of this record. Such a piece is called CRISPR RNA (crRNA). These Cas9 proteins search for DNA complementary with the crRNA. Once found, the protein breaks down or '*cleaves*' this DNA to disable it [7].

However, the intruding phages have found a way to get around this security system: by evolution their DNA changes regularly, which makes the record of the CRISPR-Cas system outdated. Therefore the Cas9 proteins must also cleave DNA which is nearly complementary to the crRNA to prevent infections from evolved phages. One could wonder when the protein decides that such strands match sufficiently and cleaves. Klein et al. [7] describe a model which explains the physics behind this decision: the model is able to foretell whether a DNA sequence is likely to be cleaved, given the crRNA. However, it cannot predict the expected time it takes to make this decision.

As techniques have developed lately, it has become possible for humans to produce crRNA ourselves, called *guide RNA* (gRNA), and embed it in a Cas9 protein such that the protein cleaves a strand of DNA of our choice [3]. This technique allows us to experiment with gene editing and in fact it is a widely used technique in the lab already [12]. This makes CRISPR-Cas9 a very promising tool for medical applications such as prevention of genetic diseases. Once it is known which sequence of DNA is responsible for such a genetic disease, scientists can produce complementary gRNA and embed it in a Cas9 protein. After that, the protein is inserted into a cell or embryo and it cleaves the desired sequence.

One might imagine that the physics behind Cas9's decision to cleave must be understood before it can be used for medical applications. While hunting for a DNA sequence which causes disease, it might cleave another sufficiently complementary DNA sequence which is responsible for vital functions. Then the patient is healed from the disease but he is lumbered with a worse malfunction.

This thesis investigates the expected time it takes the CRISPR-Cas system to cleave or not to cleave, assuming the model described in [7]. It does so by deriving expressions for general birth and death processes and applying them to the model. First the model is described in section 2. Then the expressions are derived using the mathematical theories of Markov chains and semigroups. This theory is described in section 3 and the derivations are done in section 4. After that, the expressions are interpreted in terms of the model in section 5. Finally, the moment generating function of the stopping time is derived for two special Markov processes; a random walk and a Brownian motion, using martingales. This is done in section 6.

[1] Prokaryotes are single-celled organisms without a membrane-bound nucleus.

[2] CRISPR is an abbreviation of Clustered Regularly Interspaced Short Palindromic Repeats.

2 Hybridization kinetics: a model for selection rules

In the lab several phenomena have been observed which seem to cause cleavage or unbinding of Cas9 proteins. Klein et al. describe a model to explain these targeting rules by physics in [7]. This section gives an introduction to a slightly modified version of this model.

In three steps, the Cas9 protein checks whether a part of a DNA sequence matches with the gRNA it possesses:

1) PAM binding. Initially, the protein is unbound from the DNA. By a random walk it searches a suitable place to bind, called a PAM (Protospacer Adjacent Motif). Only when it is bound to a PAM, a Cas9 protein is able to rip the DNA-helix and start the next step. These PAMs are spread throughout the DNA and they are a sequence of 3-4 nucleotides[3] (nt) [1].

2) R-loop. After being bound, Cas9 starts what is called an *R-loop*. In this R-loop it compares the DNA with the gRNA it possesses. For that purpose it separates the double helix of the DNA and tries to bind the nucleotides of the gRNA to it one by one, starting with the nucleotides at the PAM (see fig. 1). When a nucleotide of the gRNA and the DNA form a correct Watson-Crick base pair[4], the two will bind and energy is rewarded. However if the two do not match, they do not fit together chemically. Therefore once two nucleotides are bound, it costs less energy to unbind the two again if they are a mismatch compared to when they are a match. It can be said that non-matching gRNA and DNA tend to unbind again.

3) Cleavage. In the end Cas9 cleaves or unbinds depending on the progress of the R-loop. If the DNA and gRNA sequences are not sufficiently complementary, the R-loop will not be finished since the gRNA and DNA tend to unbind. Then Cas9 unbinds from the DNA and will try to find a new PAM to bind to. In contrast, if the protein is able to finish the R-loop, the two sequences are sufficiently complementary. Then Cas9 is able to cleave the DNA.

[3]DNA and RNA are built up of a sequence of four different *nucleotides*, labelled by the letters A, C, G, T and A, C, G, U respectively.

[4]The different nucleotides of DNA and RNA can make certain pairs only: C binds to G only and A binds to T or U only, and vice versa. These combinations are called *Watson-Crick base pairs*.

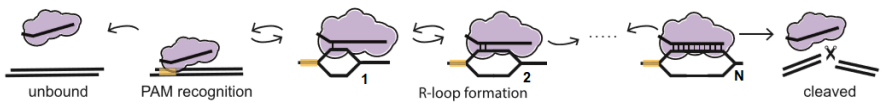


Figure 1: A sketch of the decision making process. First the Cas9 protein to the PAM. Then it starts an R-loop by binding the nucleotides one by one. If the DNA and gRNA are sufficiently complementary the DNA strand can be cleaved, else the protein will unbind. [7]

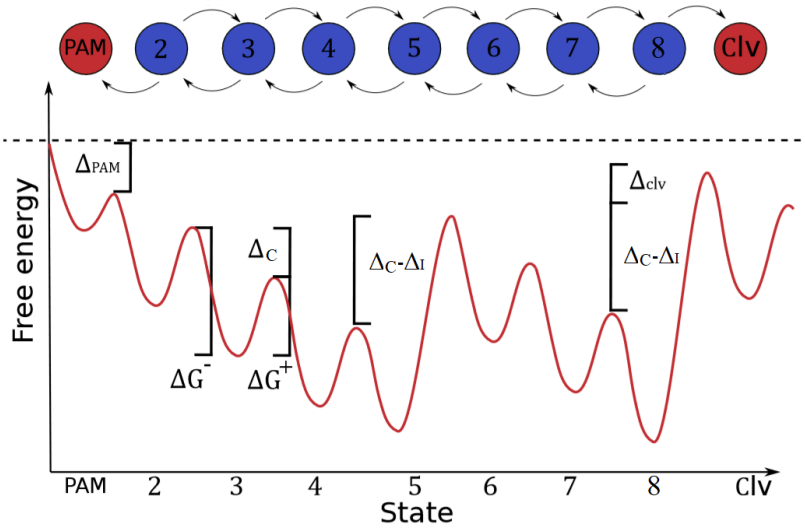


Figure 2: An example of an energy landscape [1]. In this example the gRNA length is 8 nt. There are mismatches at positions 5 and 8.
In general the system drifts towards cleavage, especially when a state matches. Therefore the barrier to the next state is Δ_C lower than the barrier to the previous state. However, the system prefers to unbind a mismatch. Therefore the free energy of a state with a mismatch is Δ_I higher than the previous one with a match.
A Cas9 protein has to bind to a PAM before it starts the R-loop. The degree of (dis-)favour to bind to such a PAM is given by Δ_{PAM}. A positive value for Δ_{PAM} implies that the Cas9 protein prefers to bind to a PAM; then barrier towards the R-loop is lower than towards unbinding.
Finally cleavage costs energy. These costs are given by Δ_{clv}.
In general, the height of the barrier of moving backwards is modelled by ΔG^- and of moving forwards by ΔG^+.
All these parameters are in units of $k_B T$.

2.1 The minimal model

As discussed, non-matching DNA and gRNA unbind more easily than a matching combination. Therefore, since a system prefers a state with the least energy, it can be said that the energy level lower when a non-matching base pair is bound than when it is unbound. With this information it is possible to draw an energy landscape for a given combination of DNA and gRNA. The energy gain or loss when Cas9 moves from one nucleotide (or *state*) to another during the R-loop can be modelled by several parameters as shown in fig. 2.

Let us define these parameters formally. First the zero energy level should be defined. The system has zero energy when the Cas9 protein is unbound from the PAM as this is its intial position.

During the R-loop the system hops from state to state using single nucleotide steps. However when it hops forward, the double helix of the DNA has to be separated, which costs energy, before the DNA and gRNA are bound, which might release some energy. On the other hand, when taking a step backward the gRNA and DNA have to be separated, which costs energy, before the DNA

helix is bound together, which releases energy. Therefore, intially, the system has to pay energy to move to another state before energy is released. This implies that there is an energy hill between every two subsequent states. Let us therefore define ΔG_i^- and ΔG_i^+, which are the height of the hill which is passed when moving to state $i-1$ and state $i+1$ respectively. They are a measure for how much free energy the system must have such that it can step to the next state.

One can imagine that the difference in the height of two barriers surrounding any state plays an important role: it tells whether the Cas protein is more probable to move to the next or to the previous state. Therefore, for any state i, the difference between these heights is given by $\Delta(i) = \Delta G_i^- - \Delta G_i^+$. The value for $\Delta(i)$ depends on the position i, however, in a minimal model, there are four different possible values for $\Delta(i)$.

First of all, the protein has to bind to a PAM before starting an R-loop. Different types of Cas proteins have a different preference or distaste to be bound to such a PAM. If it prefers to be bound, unbinding from the PAM costs more energy than starting an R-loop. This implies a difference in height between the two energy barriers surrounding the PAM state. This energy difference is given by the parameter Δ_{PAM} and it gives the degree of preference to bind to a PAM. A positive value implies a preference for binding, a negative value implies that it rather unbinds. In this thesis, only positive values of Δ_{PAM} are considered.

After that the system runs through the R-loop. In general, the system drifts towards cleavage. Therefore, in general, the barrier to a state forward is lower than the barrier to a state backwards. This energy difference is given by the parameter Δ_{C}. However when there is an off-target, (a mismatch), the system prefers to unbind such combination. Therefore, in case of a mismatch, the barrier to the next state is higher than the barrier to the previous state. The appearing energy barrier is given by the parameter Δ_{I}. Due to the drift towards cleavage, it follows that the barrier due to a mismatch is $\Delta_{\mathrm{C}} - \Delta_{\mathrm{I}}$ higher than its previous barrier.

Finally, at the last state, it costs energy to cleave the strand of DNA. This energy cost is given by Δ_{clv}. The complementarity of the last base pair plays a role now. When the last base pair is a mismatch, the energy cost for cleavage is $\Delta(N-1) = \Delta_{\mathrm{C}} - \Delta_{\mathrm{I}} - \Delta_{\mathrm{clv}}$. Else the energy cost is $\Delta(N-1) = \Delta_{\mathrm{C}} - \Delta_{\mathrm{clv}}$.

2.2 Decision rule

Having defined all parameters of the model, the decision rule of the CRISPR-Cas system can be introduced. The model of Klein et al. [7] gives the following physical rule to decide whether the Cas9 probably cleaves or unbinds:

> The CRISPR associated protein is more likely to unbind if the highest energy barrier is higher than the initial energy. On the other hand, it is more likely to cleave if the highest barrier is lower than the initial energy.

An example is shown in fig. 3. Three phenomena that have been observed in the lab cause us to observe this rule [7]:

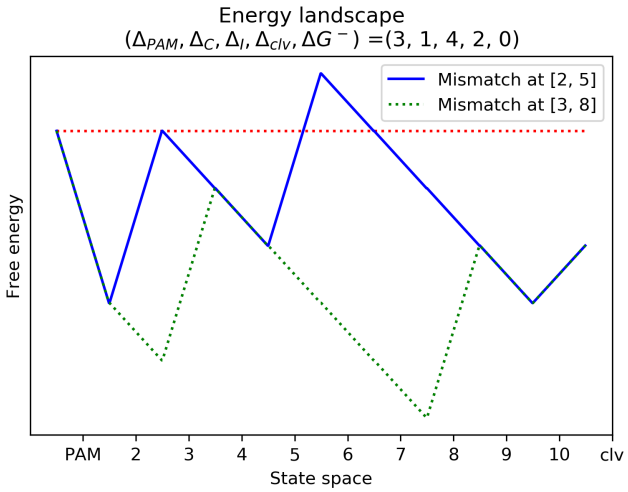


Figure 3: Two energy landscapes displayed. Cas9 is more likely to cleave in the green, dotted energy landscape than in the blue energy landscape since its highest energy barrier is below the initial free energy level.

i) Seed region. First of all lab experiments show that there is a specific region in which an off-target almost surely implies that Cas9 unbinds. This region is called the *seed region*. The last state in this seed region is given by n_{seed} and it can be calculated by

$$n_{\text{seed}} = 1 + \frac{\Delta_{\text{I}} - \Delta_{\text{PAM}}}{\Delta_{\text{C}}}. \tag{2.1}$$

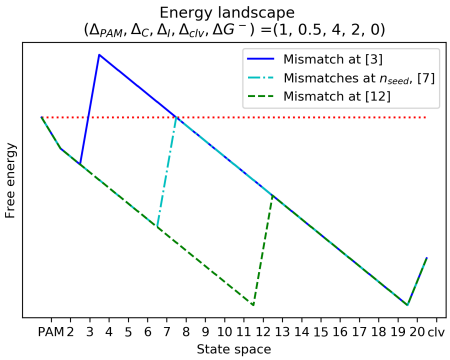
This seed region can be explained using the model developed in [7]. If there is an off-target close to the PAM, the system has not gained enough free energy to compensate for the energy penalty. Therefore an energy barrier appears which is higher than the initial energy of the system. Due to the high energy costs towards cleavage, it will cause Cas9 to unbind.

However if the first off-target is outside the seed region, the system has gained some free energy already by binding the preceding matching base pairs. Therefore the energy barrier caused by the mismatch can be compensated by this free energy. This means that the barrier is not higher than the barrier to be passed for unbinding, so the system is likely to cleave (see fig. 4a).

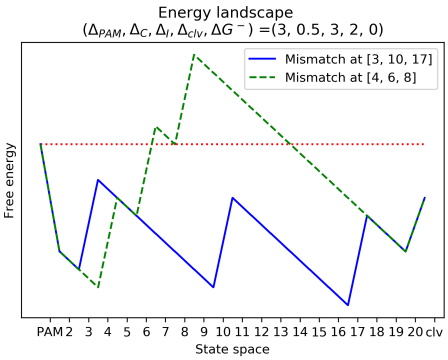
ii) Mismatch spread. Secondly it appears that a block of mismatches reduces the probability of cleavage significantly, compared to the same number of mismatches being spread throughout the DNA. This can be explained by the model. A block of mismatches causes multiple subsequent energy penalties. Therefore a very high barrier appears in the energy landscape. Such a high energy barrier makes that the system prefers unbinding above cleavage.

However, when the mismatches are spread, several lower energy barriers appear. These individual barriers are not higher than the barrier towards unbinding. Therefore the system cleaves. (see fig. 4b)

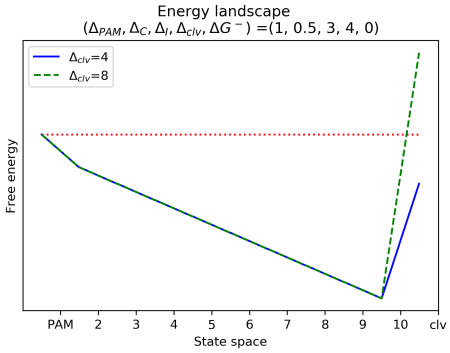
iii) Cleavage costs. The energy cost to cleave a strand of DNA is modelled by the parameter Δ_{clv}. If cleavage costs a lot of energy, the barrier towards cleavage becomes higher than the barrier to unbinding. Therefore such a system has a higher probability of unbinding. This phenomenon can be observed in fig. 4c.



(a) Three energy landscapes with different locations of a mismatch. The blue line has a mismatch before n_{seed}, the green, dashed, dotted one at n_{seed}, the dashed cyan one behind n_{seed}. Cas9 is very likely to unbind in the blue landscape while it will probably cleave in the cyan landscape.



(b) Two energy landscapes with three mismatches. A block of mismatches appears in the green, dashed landscape, while they are spread out in the blue line. The protein is very likely to unbind in the green energy landscape, while it will probably cleave in the blue energy landscape.



(c) Two energy landscapes with different values for Δ_{clv}. Cas9 is likely to cleave in the blue energy landscape as the cleavage barrier is lower than the barrier towards unbinding. The protein is more likely to unbind in the green, dashed energy landscape due to the high energy costs of cleavage.

Figure 4

2.3 CRISPR-Cas as a birth and death process

Klein et al [7] are also able to describe the cleavage process mathematically: the system can be seen as a birth and death process. In such a process, one should see the R-loop as a discrete and finite state space $\mathbb{S} := \{0, 1, \ldots, N-1, N\}$ with a walker (the Cas9 protein) on it which hops through the space continuously in time. This walker steps to the right with rate p_x and to the left with rate q_x, until it reaches the one of the boundaries of $\mathbb{S}$: then the process stops. These boundaries 0 and N are called *absorbers* and there $q_N = p_0 = 0$. Physically one could say that position 0 is the unbound state, position 1 is the PAM state and that the DNA is cleaved at position N. The state space is displayed in fig. 5

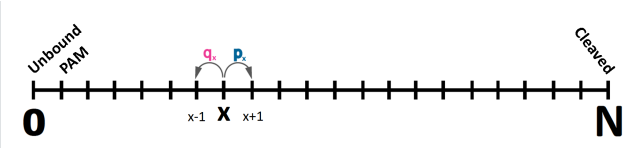


Figure 5: The state space of a birth and death process with which the system can be described. States 0 and N are absorbers and are the unbound and cleaved state respectively. State 1 is the PAM and therefore the initial positions of the walkers. They move left and right with position dependent rates q_x and p_x respectively.

The rates p_x and q_x can be related to the five parameters as defined in fig. 2 using the Arrhenius equation [1]. According to these equations, the rates are of the following form for a certain $k_0 \in \mathbb{R}^+$:

$$p_x = k_0 e^{-\Delta G_x^+} = k_0 e^{-\left(\Delta G_x^- - \Delta(x)\right)} \tag{2.2}$$

$$q_x = k_0 e^{-\Delta G_x^-} \tag{2.3}$$

In the minimal model, $\Delta(x)$ can take four values which can be deduced from fig. 2. For $x = 1$, $\Delta(1) = \Delta_{\mathrm{PAM}}$. For the other states $\Delta(x) = \Delta_{\mathrm{C}}$, however when there is a mismatch at state x, $\Delta(x) = \Delta_{\mathrm{C}} - \Delta_{\mathrm{I}}$. Finally at $x = N-1$, $\Delta(N-1) = \Delta_{\mathrm{C}} - \Delta_{\mathrm{clv}}$ or $\Delta(N-1) = \Delta_{\mathrm{C}} - \Delta_{\mathrm{I}} - \Delta_{\mathrm{clv}}$, depending on the presence of a mismatch at the last state. By taking ΔG^-, and therefore q_x constant, the rates reduce to the following:

$$p_x = \begin{cases} k_0 e^{-(\Delta G^- - \Delta_{\mathrm{PAM}})} & \text{for } x = 1 \\ k_0 e^{-(\Delta G^- - \Delta_{\mathrm{C}})} & \text{for a state with a match} \\ k_0 e^{-(\Delta G^- - \Delta_{\mathrm{C}} + \Delta_{\mathrm{I}})} & \text{for a state with a mismatch} \\ k_0 e^{-(\Delta G^- - \Delta_{\mathrm{C}} + \Delta_{\mathrm{clv}})} & \text{for } x = N-1 \text{ with a match} \\ k_0 e^{-(\Delta G^- - \Delta_{\mathrm{C}} + \Delta_{\mathrm{I}} + \Delta_{\mathrm{clv}})} & \text{for } x = N-1 \text{ with a mismatch} \end{cases} \tag{2.4}$$

$$q_x = k_0 e^{-\Delta G^-} \tag{2.5}$$

The above expressions make it possible to rewrite expressions for general birth and death processes into expressions suitable for the energy landscape of the CRISPR-Cas system.

Notice that a general birth and death processes allows the walkers to start anywhere on the state space. In the setting of CRISPR-Cas, however, the only

start position is $x = 1$ as Cas9 always starts from the PAM. Therefore we are specifically interested in expressions for birth and death processes that take $x = 1$ as the initial position of the walkers.

In the lab however it sometimes is interesting to have the Cas9 protein start at position $x = N - 1$, the state before cleavage. However, in that case the state space can be mirrored for the calculations. Then the walker starts at $x = 1$ again and the aforementioned expressions can be used.

3 Markov chains and semigroups

In section 4, expressions for birth and death processes will be derived using the theory of Markov chains and semigroups. Therefore, an introduction to this theory is given in this section. The aim is to derive the generator of a birth and death process, which will be used frequently in the next section.

Let us consider a discrete, finite, one-dimensional state space $\mathbb{S} = \{0, 1, \ldots, N\}$ and let $\{X_t, t \geq 0\}$ be a continuous-time Markov process on $\mathbb{S}$. At every position x the process moves to position y in time t with probability $p_{x,y}(t) := \mathbb{P}[X_t = y | X_0 = x]$. In this section several properties of such a process are discovered.

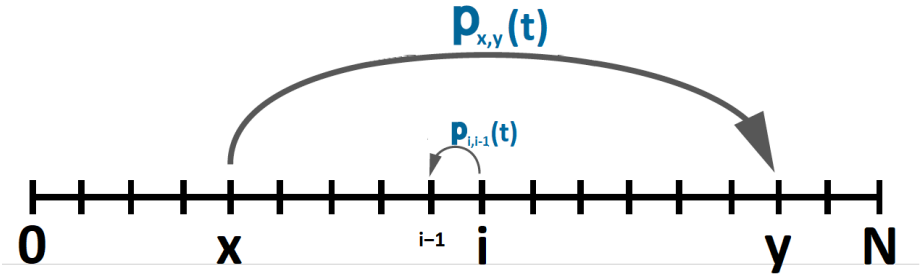


Figure 6: The discrete and finite state space $\mathbb{S}$ of a general, continuous-time Markov process. For every $x, y \in \mathbb{S}$, the probability to move from x to y within time t is given by $p_{x,y}(t)$.

A similar way to describe the probability to move from x to y is by using rates. At every position the process moves from position x to y with rate $r_{x,y}$. These rates are related to the probabilities by

$$r_{i,j} = \lim_{t \to 0} \frac{p_{i,j}(t)}{t}. \tag{3.1}$$

First a Markov process needs to be defined. Several definitions are used in literature and the following will be used here [11]

Definition 3.1 (Markov)**.** *A process $\{X_t\}$ on $\mathbb{S}$ is said to be Markovian if $\forall k \in \mathbb{N}, \forall s_0 < s_1 < \ldots < s_{k-1} < t$,*

$$\mathbb{P}(X_t = j | X_{s_0} = i_0, X_{s_1} = i_1, \ldots, X_{s_{k-1}} = i_{k-1}) = \mathbb{P}(X_t = j | X_{s_{k-1}} = i_{k-1})$$

For a Markov process, given the present, future and past are independent. Due to this property the transition matrix S_t of a Markov process can be defined. It displays the probability with which the process moves from a position $x \in \mathbb{S}$ to a position $y \in \mathbb{S}$ within a time t:

$$S_t = \begin{bmatrix} p_{0,0}(t) & p_{0,1}(t) & \cdots & p_{0,N}(t) \\ p_{1,0}(t) & p_{1,1}(t) & \cdots & p_{1,N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ p_{N,0}(t) & p_{N,1}(t) & \cdots & p_{N,N}(t) \end{bmatrix}.$$

This matrix is called the *(transition) semigroup* of the Markov process $\{X_t\}$ [9].

It is possible to define a function $f : \mathbb{S} \to \mathbb{R}$ on the state space; such a function assigns a value to every possible position of the walker in $\mathbb{S}$. Since

discrete and finite state spaces are considered only, a finite number of positions needs to be assigned to a value. That is why such a function f can be identified with the column vector

$$f = \begin{bmatrix} f(0) \\ f(1) \\ \vdots \\ f(N) \end{bmatrix} = \big[f(j)\big]_{j\in\mathbb{S}}.$$

Furthermore let us introduce the following notation:

$$\mathbb{P}_x(.) = \mathbb{P}[.|X_0 = x]$$
$$\mathbb{E}_x(.) = \mathbb{E}[.|X_0 = x]$$

Proposition 3.2. *For a function $f : \mathbb{S} \to \mathbb{R}$ defined on the state space of a Markov process, the expectation of the function can be calculated as follows:*

$$\mathbb{E}_x[f(X_t)] = (S_t f)(x)$$

Proof. Since f is defined on $\mathbb{S}$ which is discrete and finite, it can be considered a column vector $[f(j)]_{j\in\mathbb{S}}$. Then $S_t f$ can be seen as a matrix-vector product, which is written as the following column vector:

$$S_t f = \left[\sum_{j\in\mathbb{S}} p_{i,j}(t) \cdot f(j)\right]_{i\in\mathbb{S}}$$

Using this result and the definition of the expectation value one can rewrite:

$$\begin{aligned} \mathbb{E}_x[f(X_t)] &= \sum_{j\in\mathbb{S}} f(j)\mathbb{P}_x[X_t = j] \\ &= \sum_{j\in\mathbb{S}} f(j)p_{x,j}(t) \\ &= S_t f(x) \end{aligned} \tag{3.2}$$

This last step follows from the aforementioned matrix-vector multiplication. □

Furthermore let us derive the following lemma [9]:

Lemma 3.3. *Given two random variables X, Y, the following holds:*

$$\mathbb{E}\big[\mathbb{E}[X|Y]\big] = \mathbb{E}[X]$$

Proof. Assume that X, Y are two random variables, then

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}[X|Y]\big] &= \sum_{k=-\infty}^{\infty} \mathbb{E}[X|Y=k]\mathbb{P}(Y=k) \\
&= \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} n\mathbb{E}[X=n|Y=k]\mathbb{P}(Y=k) \\
&= \sum_{n=-\infty}^{\infty} n \sum_{k=-\infty}^{\infty} \mathbb{P}[X=n|Y=k]\mathbb{P}(Y=k) \\
&= \sum_{n=-\infty}^{\infty} n \sum_{k=-\infty}^{\infty} \mathbb{P}(X=n \cap Y=k) \\
&\overset{*}{=} \sum_{n=-\infty}^{\infty} n\mathbb{P}(X=n) \\
&= \mathbb{E}[X]
\end{aligned}$$

In the step marked with $*$ the following property was used:

$$\mathbb{P}(X=n) = \sum_{k=-\infty}^{\infty} \mathbb{P}(X=n \cap Y=k).$$

This is true by the law of total probability. □

Now a useful property of semigroups can be derived: the *semigroup property* [11]:

Proposition 3.4 (Semigroup property)**.** *For any points in time $s, t \geq 0$*

$$S_{t+s} = S_t S_s. \tag{3.3}$$

As a consequence $S_0 = I$.

Proof. Using proposition 3.2 one finds

$$S_{t+s}f(x) = \mathbb{E}_x[f(X_{t+s})]$$

Then by lemma 3.3:

$$= \mathbb{E}_x\big[\mathbb{E}[f(X_{t+s})|X_s]\big]$$

The Markov property allows us to make a time shift and set $X_0 = X_s$:

$$= \mathbb{E}_x\big[\mathbb{E}_{X_s}[f(X_t)]\big]$$

Using proposition 3.2 twice one finds

$$\begin{aligned}
&= \mathbb{E}_x[S_t f(X_s)] \\
&= (S_s(S_t f))(x) \\
&= S_t S_s f(x)
\end{aligned}$$

This last step follows because the role of s and t can be reversed to obtain $S_{t+s}f = S_t S_s f$.

Notice that the following follows from the semigroup property: $IS_t = S_t = S_{0+t} = S_0 S_t$. Therefore $S_0 = I$. □

From the semigroup property, one can also derive the Chapman-Kolmogorov equations:

$$p_{i,j}(t+s) = \sum_{k \in \mathbb{S}} p_{i,k}(t) p_{k,j}(s). \tag{3.4}$$

This equation states that the process moves from states i to j in time $t+s$ by moving from state i to any state k in time t and then moving from k to j in the remaining time s.

Given the semigroup property, the semigroup can be expressed in terms of a generator. Differentiating the semigroup with respect to t:

$$\begin{aligned} S_t' &= \lim_{h \to 0} \frac{S_{t+h} - S_t}{h} \\ &= \lim_{h \to 0} \frac{S_t S_h - S_t}{h} \\ &= S_t \lim_{h \to 0} \frac{S_h - I}{h} \\ &= S_t L \end{aligned} \tag{3.5}$$

Where $L := \lim_{h \to 0} \frac{S_h - I}{h}$. This matrix L is called the *generator* of the Markov process. The differential equation in eq. (3.5) gives the following result:

$$S_t = S_0 e^{tL} = e^{tL}. \tag{3.6}$$

Assuming $t \ll 1$, this expression for S_t can be simplified by the definition of the matrix exponential:

$$\begin{aligned} S_t f(x) = e^{tL} f(x) &= \sum_{n=0}^{\infty} \frac{1}{n!} (tL)^n f(x) \\ &= f(x) + tLf(x) + O(t^2) \end{aligned} \tag{3.7}$$

3.1 Birth and death processes

Semigroups can be applied to birth and death processes as they are a special type of Markov processes. As defined in section 2.3, a birth and death process $\{X_t, t \geq 0\}$ is a continuous-time process defined on a discrete state space. At every position x the process moves right with rate p_x and left with rate q_x. For $t \ll 1$ the walker will make one step at most with probability close to 1. Therefore within a time t, the process has either taken one step to the right, one to the left or it has not moved yet. The probabilities of these options are

$$\begin{aligned} \mathbb{P}_x(X_t = x+1) &= tp_x + O(t^2) \\ \mathbb{P}_x(X_t = x-1) &= tq_x + O(t^2) \\ \mathbb{P}_x(X_t = x) &= 1 - tp_x - tq_x + O(t^2) \end{aligned}$$

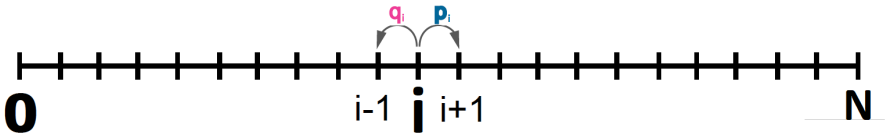


Figure 7: A scheme of the state space of a birth and death process. A finite, discrete state space $\mathbb{S}$ is considered. At every position i, the walkers move left with rate q_i and right with rate p_i.

Then for a function $f : \mathbb{S} \to \mathbb{R}$ the expected value for $f(X_t)$ can be deduced:

$$\begin{aligned}\mathbb{E}_x[f(X_t)] &= \sum_{j\in\mathbb{S}} \mathbb{P}_x(X_t = j) f(j) \qquad (3.8)\\ &= tp_x f(x+1) + tq_x f(x-1) + (1 - tp_x - tq_x) f(x) + O(t^2)\\ &= f(x) + t\Big(p_x\big(f(x+1) - f(x)\big) + q_x\big(f(x-1) - f(x)\big)\Big) + O(t^2) \qquad (3.9)\end{aligned}$$

Note that according to proposition 3.2 this result is equal to $S_t f(x)$. Therefore this expression can be seen as the semigroup of a birth and death process. By the definition of a generator, an expression for the generator of a birth and death process can be found:

$$\begin{aligned}Lf(x) &= \lim_{t\to 0} \frac{S_t f(x) - f(x)}{t}\\ &= \lim_{t\to 0} \big\{p_x\big(f(x+1) - f(x)\big) + q_x\big(f(x-1) - f(x)\big) + O(t)\big\}\\ &= p_x\big(f(x+1) - f(x)\big) + q_x\big(f(x-1) - f(x)\big). \qquad (3.10)\end{aligned}$$

This generator plays an important role in the derivations in the next chapter.

4 Expressions for stopping times of birth and death processes

Recall the birth and death process defined in section 2.3. A finite state space $\mathbb{S} = \{0, 1, 2, \ldots, N\}$ is considered and $\{X_t, t \geq 0\}$ is a birth and death process defined on $\mathbb{S}$. For every $x \in \mathbb{S}$, the process moves to the right with rate p_x and to the left with q_x. Absorbers can be found at positions $x = 0$ and $x = N$, which are physically interpreted as the unbound state and the cleavage state. Furthermore let us define the following:

$$T_0 = \inf\{t \geq 0 : X_t = 0\} \text{ i.e. the unbinding time.}$$
$$T_N = \inf\{t \geq 0 : X_t = N\} \text{ i.e. the cleavage time.}$$
$$T_{0,N} = \inf\left\{t \geq 0 : X_t \in \{0, N\}\right\} \text{ i.e. the time to cleave or unbind.}$$

In this section closed expressions for the following functions are derived:

1. $\mathcal{P}_{\text{clv}}(x) := \mathbb{P}_x(T_N < T_0)$. The probability of arriving at N before arriving at 0, starting from x;
2. $\mathcal{T}(x) := \mathbb{E}_x(T_{0,N})$. The expected time taken to arrive at either position 0 or N, starting from x;
3. $\mathcal{T}_{\text{ub}}(x) := \mathbb{E}_x(T_{0,N}|T_0 < T_N)$. The expected time taken to arrive at 0 before arriving at N, starting from x;
4. $\mathcal{T}_{\text{clv}}(x) := \mathbb{E}_x(T_{0,N}|T_N < T_0)$. The expected time taken to arrive at N before arriving at 0, starting from x.

First the following lemmas will be proved:

Lemma 4.1. *For every function $f(x)$ the following equivalence holds:*

(a) f is harmonic, *i.e. $S_t f = f$, $\forall t \geq 0$*

(b) Lf=0

Proof. Starting by proving (a)⇒(b), it is assumed that $S_t f = f$ one can find:

$$S_t f = f \Rightarrow (S_t - I)f = 0 \Rightarrow \frac{S_t - I}{t} f = 0$$

Then in the limit of $t \to 0$ the following is found by the definition of the generator L:

$$\lim_{t \to 0} \frac{S_t - I}{t} f = Lf = 0$$

In proving (b) ⇒ (a), $Lf = 0$ can be rewritten as follows by multiplying both sides by S_t and using eq. (3.5):

$$S_t L f = \frac{d}{dt} S_t f = 0.$$

This implies that $S_t f$ is independent of time. Therefore, using that $S_0 = I$, the following can be written:

$$S_t f = S_0 f = f.$$

Now the equivalence has been proved. □

Lemma 4.2. *For all functions $f(x), g(x)$, the following implication holds:*

$$\forall t \geq 0,\ S_t f - f = tg \Longrightarrow Lf = g$$

Proof. Let us assume $S_t f - f = tg$. This is equivalent to $\frac{S_t - I}{t} f = g$. Since f and g are functions of x, taking the limit $t \downarrow 0$ gives

$$\lim_{t \downarrow 0} \frac{S_t - I}{t} f(x) = Lf(x) = g(x)$$

which proves the implication. □

Finally the following definition plays an important role in the desired expressions:

Definition 4.3. *The function $\varphi(x)$ is defined as*

$$\varphi(x) = \begin{cases} \prod_{j=1}^{x} \frac{q_j}{p_j} & \text{if } x \geq 1 \\ 1 & \text{if } x = 0 \end{cases}$$

4.1 Probability to cleave

In this subsection a closed expression for the probability to cleave, $\mathcal{P}_{\text{clv}}(x)$, is derived. Recall the definition $\mathcal{P}_{\text{clv}}(x) := \mathbb{P}_x(T_N < T_0)$. Using the partition theorem, this can be written as:

$$\mathcal{P}_{\text{clv}}(x) = \sum_{y \in \mathbb{S}} \mathbb{P}_x(T_N < T_0 | X_t = y) \mathbb{P}_x(X_t = y)$$

By the Markov property, the information $\{X_s : s < t\}$ can be ignored. Therefore one can say that $X_0 = y$:

$$\begin{aligned} &= \sum_{y \in \mathbb{S}} \mathbb{P}_y(T_N < T_0) \mathbb{P}_x(X_t = y) \\ &= \sum_{y \in \mathbb{S}} \mathcal{P}_{\text{clv}}(y) \mathbb{P}_x(X_t = y). \end{aligned}$$

Then according to eq. (3.2) we find

$$\mathcal{P}_{\text{clv}}(x) = S_t \mathcal{P}_{\text{clv}}(x)$$

One can conclude that $\mathcal{P}_{\text{clv}}$ is harmonic, hence, according to lemma 4.1

$$L\mathcal{P}_{\text{clv}} = 0. \tag{4.1}$$

This is a difference equation for $\mathcal{P}_{\text{clv}}(x)$ which can be solved using the expression for L given in eq. (3.9). This is done in appendix A.1 and the result is

$$\mathcal{P}_{\text{clv}}(y) = \frac{\sum_{x=0}^{y-1} \varphi(x)}{\sum_{x=0}^{N-1} \varphi(x)} \tag{4.2}$$

with y the starting position of the walker and $\varphi(x)$ given in definition 4.3.

As described in section 2.3, the only start position of interest is $y = 1$. Then this expression reduces to

$$\mathcal{P}_{\text{clv}}(1) = \frac{1}{\sum_{x=0}^{N-1} \varphi(x)}. \tag{4.3}$$

Note that this expression coincides with the expression found by Klein et al. in [7] with other methods.

Probability in terms of deltas

Now the derived expression can be rewritten in terms of the deltas described in section 2. First the expression for $\varphi(x)$ is rewritten. Recall its definition in definition 4.3 and that $\Delta(i) = \Delta G_i^- - \Delta G_i^+$ for every position i in the state space. Using these definitions and eqs. (2.2) and (2.3), the following can be seen:

$$\varphi(x) = \prod_{j=1}^{x} \frac{q_j}{p_j} = \prod_{j=1}^{x} e^{-(\Delta G_j^- - \Delta G_j^+)} = e^{-\sum_{j=1}^{x} \Delta(j)} \tag{4.4}$$

Then

$$\mathcal{P}_{\text{clv}}(y) = \frac{1 + \sum_{x=1}^{y-1} e^{-\sum_{j=1}^{x} \Delta(j)}}{1 + \sum_{x=1}^{N-1} e^{-\sum_{j=1}^{x} \Delta(j)}} \tag{4.5}$$

and

$$\mathcal{P}_{\text{clv}}(1) = \left(\sum_{x=0}^{N-1} \varphi(x) \right)^{-1} = \left(1 + \sum_{x=1}^{N-1} e^{-\sum_{j=1}^{x} \Delta(j)} \right)^{-1}. \tag{4.6}$$

As described in section 2.3, $\Delta(j)$ can take five values in the minimal model:

$$\Delta(j) = \begin{cases} \Delta_{\text{PAM}} & \text{for } j = 1 \\ \Delta_{\text{C}} & \text{for a state } j \notin \{1, N-1\} \text{ with a match} \\ \Delta_{\text{C}} - \Delta_{\text{I}} & \text{for a state } j \notin \{1, N-1\} \text{ with a mismatch} \\ \Delta_{\text{C}} - \Delta_{\text{clv}} & \text{for } j = N-1 \text{ with a match} \\ \Delta_{\text{C}} - \Delta_{\text{I}} - \Delta_{\text{clv}} & \text{for } j = N-1 \text{ with a mismatch} \end{cases}$$

Using this information, the sum in the exponent of eq. (4.6) can be rewritten into

$$\sum_{j=1}^{x} \Delta(j) = \Delta_{\text{PAM}} + (x-1)\Delta_{\text{C}} - c_x \Delta_{\text{I}} - \delta_{x,N-1}\Delta_{\text{clv}}. \tag{4.7}$$

In this expression $\delta_{i,j}$ is the Kronecker delta, defined as

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

and c_x is a count function: it counts the number of mismatches between states 1 and x. Notice that $(x-1)$ appears in front of Δ_{C} because all states contain the parameter Δ_{C}, except the PAM state. Combining eq. (4.7) with eqs. (4.5) and (4.6) gives an expression for $\mathcal{P}_{\text{clv}}$ and $\mathcal{P}_{\text{clv}}(1)$ respectively for the minimal model.

4.2 Expected time to cleave or unbind

In this section a closed expression for the expected time to either unbind or cleave, $\mathcal{T}(x)$, is derived. Recall the definition $\mathcal{T}(x) := \mathbb{E}_x(T_{0,N})$. Using the partition theorem, this can be written as follows:

$$\mathcal{T}(x) = \sum_{y\in\mathbb{S}} \mathbb{E}_x(T_{0,N}|X_t = y)\mathbb{P}_x(X_t = y)$$

By the Markov property, the past can be forgotten and the expectation can be shifted by a time t. However, it should not be forgotten that the walker has already run a time t. That is why t is added to $T_{0,N}$.

$$= \sum_{y\in\mathbb{S}} \mathbb{E}_y(t + T_{0,N})\mathbb{P}_x(X_t = y)$$

Since the expectation operator is linear and $t\sum_{y\in\mathbb{S}} \mathbb{P}_x(X_t = y) = t$ (the walker must be somewhere in the state space) one finds

$$\begin{aligned} &= t + \sum_{y\in\mathbb{S}} \mathbb{E}_y(T_{0,N})\mathbb{P}_x(X_t = y) \\ &= t + \sum_{y\in\mathbb{S}} \mathcal{T}(y)\mathbb{P}_x(X_t = y) \end{aligned}$$

Using eq. (3.2) the following equation is found.

$$\mathcal{T}(x) = t + S_t\mathcal{T}(x)$$

This can be rewritten as

$$S_t\mathcal{T} - \mathcal{T} = -t.$$

Now use lemma 4.2 to obtain a difference equation for $\mathcal{T}(x)$:

$$L\mathcal{T}(x) = -1. \tag{4.8}$$

This difference equation can be solved using eq. (3.10) and this is done in appendix A.2. The result is given here:

$$\mathcal{T}(y) = \sum_{x=0}^{y-1} \left\{ \frac{\sum_{\xi=0}^{N-1}\sum_{j=0}^{\xi-1} \frac{1}{p_{\xi-j}}\frac{\varphi(\xi)}{\varphi(\xi-j)}}{\sum_{\xi=0}^{N-1}\varphi(\xi)}\varphi(x) - \sum_{i=0}^{x-1} \frac{1}{p_{x-i}}\frac{\varphi(x)}{\varphi(x-i)} \right\}. \tag{4.9}$$

For initial position $y = 1$, this expression can be reduced to:

$$\mathcal{T}(1) = \frac{\sum_{\xi=0}^{N-1}\sum_{j=0}^{\xi-1} \frac{1}{p_{\xi-j}}\frac{\varphi(\xi)}{\varphi(\xi-j)}}{\sum_{\xi=0}^{N-1}\varphi(\xi)} \tag{4.10}$$

Stopping time in terms of deltas

The expressions for $\mathcal{T}(1)$ can be rewritten in the terms of the model as well. First observe the following, using eq. (4.4):

$$\begin{aligned} \frac{1}{p_{x-i}}\frac{\varphi(x)}{\varphi(x-i)} &= \frac{1}{q_{x-i}}\frac{\varphi(x)}{\varphi(x-i-1)} = \frac{1}{q_{x-i}}e^{-\left(\sum_{j=1}^{x}\Delta(j) - \sum_{j=1}^{x-i-1}\Delta(j)\right)} \\ &= \frac{1}{q_{x-i}}e^{-\sum_{j=x-i}^{x}\Delta(j)} \end{aligned} \tag{4.11}$$

Furthermore notice that the denominator of $\mathcal{T}(1)$ is equivalent to $1/\mathcal{P}_{\mathrm{clv}}(1)$ according to eq. (4.3). Using these two equivalences the following expression can be found:

$$\mathcal{T}(1) = \mathcal{P}_{\mathrm{clv}}(1) \sum_{x=1}^{N-1} \sum_{i=0}^{x-1} \frac{1}{q_{x-i}} e^{-\sum_{j=x-i}^{x} \Delta(j)}.$$

By substituting $\ell = x - i$, the expression can be written as

$$\mathcal{T}(1) = \mathcal{P}_{\mathrm{clv}}(1) \sum_{x=1}^{N-1} \sum_{\ell=1}^{x} \frac{1}{q_{\ell}} e^{-\sum_{j=\ell}^{x} \Delta(j)}.$$

Since $q_\ell = q$ is constant in the minimal model, an expression for the minimal model is:

$$\mathcal{T}(1) = \frac{\mathcal{P}_{\mathrm{clv}}(1)}{q} \sum_{x=1}^{N-1} \sum_{\ell=1}^{x} e^{-\sum_{j=\ell}^{x} \Delta(j)} \tag{4.12}$$

with

$$\sum_{j=\ell}^{x} \Delta(j) = \delta_{\ell,1}(\Delta_{\mathrm{PAM}} - \Delta_{\mathrm{C}}) + (x - \ell + 1)\Delta_{\mathrm{C}} - c_{\ell,x}\Delta_{\mathrm{I}} - \delta_{x,N-1}\Delta_{\mathrm{clv}}. \tag{4.13}$$

In this expression $\delta_{i,j}$ is the Kronecker delta and $c_{\ell,x}$ is a function which counts the number of mismatches between states ℓ and x. Notice that if the PAM state is in the range $[\ell, x]$, one value of Δ_{C} has to be subtracted since $\Delta(1)$ does not contain the parameter Δ_{C}. This explains the first term of the sum.

4.3 Expected time to cleave

A closed expressions for $\mathcal{T}_{\text{clv}}(x)$ is derived in this section. The following three quantities are defined:

1. $\mathcal{T}_{\text{clv}}(x) := \mathbb{E}_x(T_{0,N}|T_N < T_0)$;
2. $\kappa(x) := \mathbb{E}_x(T_{0,N}I(T_N < T_0))$;
3. $\mathcal{P}_{\text{clv}}(x) := \mathbb{P}_x(T_N < T_0)$

in which $I(.)$ is the indicator function. The following identity is used to derive an expression for $\mathcal{T}_{\text{clv}}(x)$:

$$\mathcal{T}_{\text{clv}}(x) = \frac{\mathbb{E}_x\left(T_{0,N}I(T_N < T_0)\right)}{\mathbb{P}_x(T_N < T_0)} = \frac{\kappa(x)}{\mathcal{P}_{\text{clv}}(x)}. \tag{4.14}$$

An expression for $\mathcal{P}_{\text{clv}}(x)$ has been derived in section 4.1, therefore an expression for $\kappa(x)$ only needs to be found. Using the partition theorem, $\kappa(x)$ can be written as follows:

$$\kappa(x) = \sum_{y\in\mathbb{S}} \mathbb{E}_x\big[T_{0,N}I(T_N < T_0)|X_t = y\big]\mathbb{P}_x(X_t = y)$$

Using the Markov property, the past can be forgotten and the time can be shifted with t:

$$= \sum_{y\in\mathbb{S}} \mathbb{E}_y\big[(t + T_{0,N})I(T_N < T_0)\big]\mathbb{P}_x(X_t = y)$$

By the linearity of the expectation operator the result can be rewritten:

$$= t\sum_{y\in\mathbb{S}} \mathbb{E}_y\big[I(T_N < T_0)\big]\mathbb{P}_x(X_t = y) + \sum_{y\in\mathbb{S}} \mathbb{E}_y\big[T_{0,N}I(T_N < T_0)\big]\mathbb{P}_x(X_t = y)$$

In the right term, the definition of $\kappa(x)$ appears. Therefore it can be substituted again. In the term on the left, the identity $\mathbb{E}\big[I(A)\big] = \mathbb{P}(A)$ is used. Then the expression reduces to

$$= t\sum_{y\in\mathbb{S}} \mathbb{P}_y(T_N < T_0)\mathbb{P}_x(X_t = y) + \sum_{y\in\mathbb{S}} \kappa(y)\mathbb{P}_x(X_t = y)$$

Finally using the partition theorem and the definition of $\mathcal{P}_{\text{clv}}$, one finds

$$= t\mathbb{P}_x(T_N < T_0) + S_t\kappa(x)$$
$$\kappa(x) = t\mathcal{P}_{\text{clv}}(x) + S_t\kappa(x)$$

This expression for $\kappa(x)$ can be rewritten such that:

$$S_t\kappa - \kappa = -t\mathcal{P}_{\text{clv}}$$

Now use lemma 4.2 to obtain

$$L\kappa(x) = -\mathcal{P}_{\text{clv}}(x). \tag{4.15}$$

This difference equation for κ is solved in appendix A.3. Combining this result for $\kappa(x)$ with eq. (4.14) one finds

$$\mathcal{T}_{\mathrm{clv}}(y) = \frac{1}{\mathcal{P}_{\mathrm{clv}}(y)} \sum_{x=0}^{y-1} \left\{ \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\mathrm{clv}}(\xi-j)}{\sum_{\xi=0}^{N-1} \varphi(\xi)} \varphi(x) - \sum_{i=0}^{x-1} \frac{\varphi(x)}{\varphi(x-i)} \frac{1}{p_{x-i}} \mathcal{P}_{\mathrm{clv}}(x-i) \right\} \tag{4.16}$$

with y the start location of the walker and $\mathcal{P}_{\mathrm{clv}}$ given by eq. (4.2).

Then for start location $y = 1$ one finds

$$\mathcal{T}_{\mathrm{clv}}(1) = \frac{1}{\mathcal{P}_{\mathrm{clv}}(1)} \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\mathrm{clv}}(\xi-j)}{\sum_{\xi=0}^{N-1} \varphi(\xi)}.$$

Notice that the denominator of this expression is equivalent to $1/\mathcal{P}_{\mathrm{clv}}(1)$. Therefore the equation reduces to

$$\mathcal{T}_{\mathrm{clv}}(1) = \sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\mathrm{clv}}(\xi-j). \tag{4.17}$$

Cleavage time in terms of deltas

Using the result obtained in eq. (4.11), one finds

$$\mathcal{T}_{\mathrm{clv}}(1) = \sum_{x=0}^{N-1} \sum_{i=0}^{x-1} \frac{1}{q_{x-i}} e^{-\sum_{j=x-i}^{x} \Delta(j)} \mathcal{P}_{\mathrm{clv}}(x-i)$$

The substitution $\ell = x - i$ gives the following expression:

$$\mathcal{T}_{\mathrm{clv}}(1) = \sum_{x=0}^{N-1} \sum_{\ell=1}^{x} \frac{1}{q_\ell} e^{-\sum_{j=\ell}^{x} \Delta(j)} \mathcal{P}_{\mathrm{clv}}(\ell) \tag{4.18}$$

In the minimal model, q_ℓ is a constant. Therefore the expression reduces to

$$\mathcal{T}_{\mathrm{clv}}(1) = \frac{1}{q} \sum_{x=0}^{N-1} \sum_{\ell=1}^{x} e^{-\sum_{j=\ell}^{x} \Delta(j)} \cdot \mathcal{P}_{\mathrm{clv}}(\ell). \tag{4.19}$$

The sum in the exponent is given by eq. (4.13) and $\mathcal{P}_{\mathrm{clv}}(\ell)$ by eq. (4.5).

4.4 Expected time to unbind

The derivation for the expected unbinding time $\mathcal{T}_{\text{ub}}$ is similar to the derivation of $\mathcal{T}_{\text{clv}}$. Some important steps in the derivation are described in this section. The following three expressions are defined:

1. $\mathcal{T}_{\text{ub}}(x) := \mathbb{E}_x(T_{0,N}|T_0 < T_N)$;
2. $\nu(x) := \mathbb{E}_x(T_{0,N}I(T_0 < T_N))$;
3. $\mathcal{P}_{\text{ub}}(x) := \mathbb{P}_x(T_0 < T_N)$.

Then the following identity can be used to derive an expression for $\mathcal{T}_{\text{ub}}$:

$$\mathcal{T}_{\text{ub}}(x) = \frac{\nu(x)}{\mathcal{P}_{\text{ub}}(x)} \tag{4.20}$$

Since there is no other option for a walker than cleavage or unbinding, it can be seen that $\mathcal{P}_{\text{ub}} = 1 - \mathcal{P}_{\text{clv}}$. Therefore an expression for $\mathcal{P}_{\text{ub}}(x)$ can be derived from the results of section 4.1 and an expression for $\nu(x)$ has to be derived only. This derivation is similar to the one of $\kappa(x)$ in the previous subsection. Then one finds that

$$S_t\nu - \nu = -t\mathcal{P}_{\text{ub}}.$$

Lemma 4.2 can be used to obtain

$$L\nu(x) = -\mathcal{P}_{\text{ub}}(x) \tag{4.21}$$

which is a difference equation for $\nu(x)$. This difference equation is solved in appendix A.4. Then using eq. (4.20) a closed expression for $\mathcal{T}_{\text{ub}}$ is found:

$$\mathcal{T}_{\text{ub}}(y) = \frac{1}{\mathcal{P}_{\text{ub}}(y)} \sum_{x=0}^{y-1} \left\{ \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\text{ub}}(\xi-j)}{\sum_{\xi=0}^{N-1} \varphi(\xi)} \varphi(x) - \sum_{i=0}^{x-1} \frac{\varphi(x)}{\varphi(x-i)} \frac{1}{p_{x-i}} \mathcal{P}_{\text{ub}}(x-i) \right\}. \tag{4.22}$$

For start location $y = 1$ this expression reduces to

$$\mathcal{T}_{\text{ub}}(1) = \frac{1}{\mathcal{P}_{\text{ub}}(1)} \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\text{ub}}(\xi-j)}{\sum_{\xi=0}^{N-1} \varphi(\xi)}. \tag{4.23}$$

Unbinding time in terms of deltas

Notice that again, the denominator of eq. (4.23) is equivalent to $1/\mathcal{P}_{\text{clv}}(1)$. In combination with eq. (4.11) and the substitution $\ell = x - i$ one finds

$$\mathcal{T}_{\text{ub}}(1) = \frac{\mathcal{P}_{\text{clv}}(1)}{\mathcal{P}_{\text{ub}}(1)} \sum_{x=0}^{N-1} \sum_{\ell=1}^{x} \frac{1}{q_\ell} e^{-\sum_{j=\ell}^{x} \Delta(j)} \mathcal{P}_{\text{ub}}(\ell). \tag{4.24}$$

Assuming that $q_\ell = q$ is constant in the minimal model, this equation reduces to

$$\mathcal{T}_{\text{ub}}(1) = \frac{\mathcal{P}_{\text{clv}}(1)}{q\mathcal{P}_{\text{ub}}(1)} \sum_{x=0}^{N-1} \sum_{\ell=1}^{x} e^{-\sum_{j=\ell}^{x} \Delta(j)} \mathcal{P}_{\text{ub}}(\ell) \tag{4.25}$$

which, in combination with eqs. (4.5) and (4.13) gives the expression for $\mathcal{T}_{\text{ub}}(1)$ in terms of the parameters of the minimal model.

5 Analysis of the expressions

In the previous chapter closed-form expressions for the following four quantities were derived:

1. $\mathcal{P}_{\text{clv}}(x) := \mathbb{P}_x(T_N < T_0)$. The probability of arriving at N before arriving at 0, starting from x;
2. $\mathcal{T}(x) := \mathbb{E}_x(T_{0,N})$. The expected time taken to arrive at either position 0 or N, starting from x;
3. $\mathcal{T}_{\text{ub}}(x) := \mathbb{E}_x(T_{0,N}|T_0 < T_N)$. The expected time taken to arrive at 0 before arriving at N, starting from x;
4. $\mathcal{T}_{\text{clv}}(x) := \mathbb{E}_x(T_{0,N}|T_N < T_0)$. The expected time taken to arrive at N before arriving at 0, starting from x.

Let us analyse these four expressions by comparing them to Gillespie simulations and interpreting them physically.

5.1 General expressions for birth and death processes

Initially, the expressions were derived for general birth and death processes. Let us compare these expressions with Gillespie simulations. For now constant rates are assumed throughout the state space which sum to 1 at every position, i.e. $p_x = p$ and $q_x = q = 1 - p$.

Consider the expressions for $\mathcal{P}_{\text{clv}}, \mathcal{T}$ and $\mathcal{T}_{\text{ub}}$ in eqs. (4.2), (4.9) and (4.22). Their values are calculated and simulated as a function of y for two different values of p. The results are displayed in fig. 8.

First of all it can be seen in figs. 8a and 8b that the exact and simulated results for $\mathcal{P}_{\text{clv}}$ perfectly overlap. Therefore, for both values of (p, q) the exact result in eq. (4.2) appears to be a very accurate expression for the probability of cleavage for any value of y.

However in figs. 8c and 8e the exact results for $\mathcal{T}$ and $\mathcal{T}_{\text{ub}}$ deviate substantially from the simulations. This deviation starts from a certain value of y, defined as y_{dev}. It can be seen in figs. 8d and 8f that y_{dev} is larger or not visible for $(p, q) = (0.4; 0.6)$, which implies a very small error. It can be hypothesised that the position of y_{dev} decreases, and therefore the error increases, for an increasing value of $(q - p)$. Multiple tests confirm this and fig. 9b shows that y_{dev} is very small for $(p, q) = (0.1; 0.9)$.

However there is no deviation for a negative value of $(q - p)$. Figure 9 shows this phenomenon by comparing the results of $(p, q) = (0.9; 0.1)$ and $(p, q) = (0.1; 0.9)$. It can be seen that the first combination does not show a difference between the exact and the simulated result, while the second combination does.

One could hypothesise that this deviation is a numerical error. Let us consider for example the expression for $\mathcal{T}$ and compare the cases $(p, q) = (0.1; 0.9)$ and $(p, q) = (0.9; 0.1)$. For constant rates p and q, eq. (4.9) can be reduced to

$$\mathcal{T}(y+1) - \mathcal{T}(y) = \underbrace{c\left(\frac{q}{p}\right)^y}_{(a)} - \underbrace{\sum_{i=0}^{y} \frac{1}{p}\left(\frac{q}{p}\right)^i}_{(b)} \tag{5.1}$$

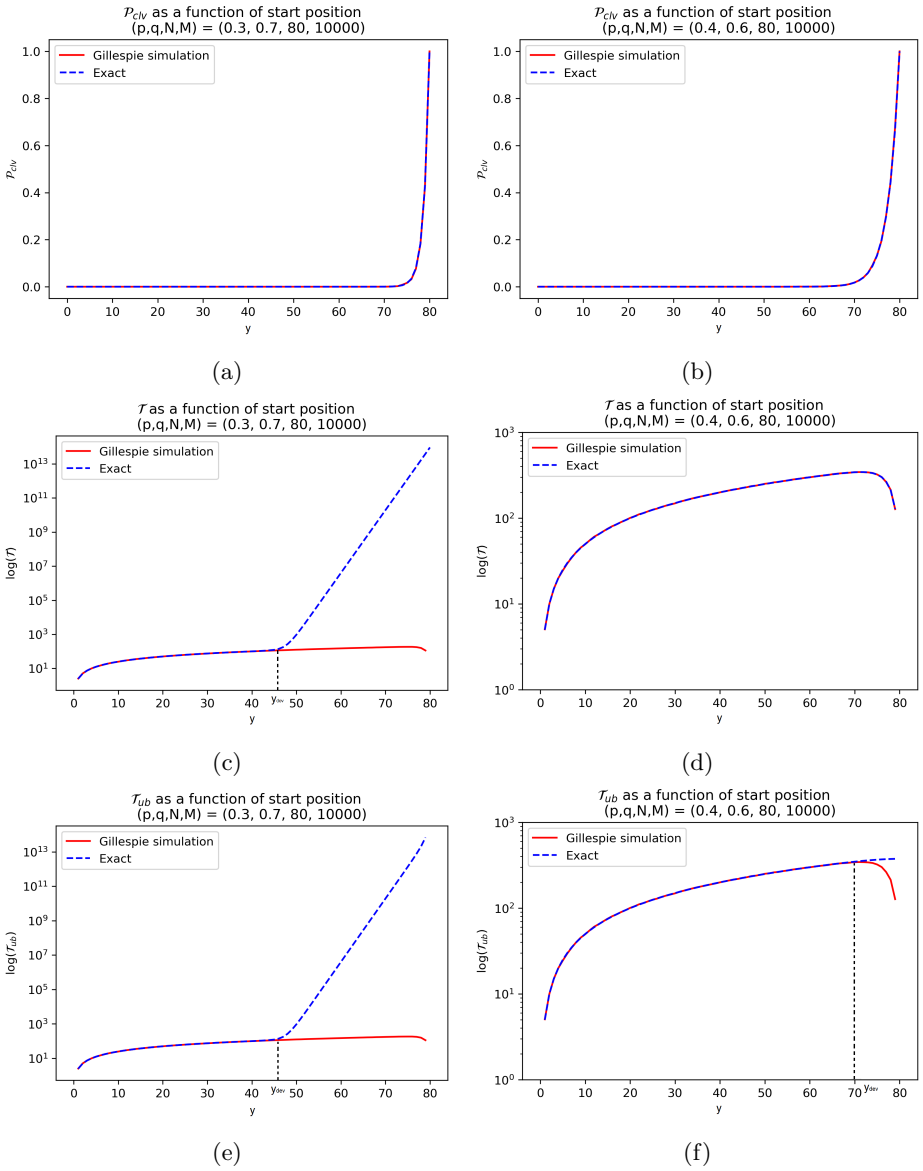
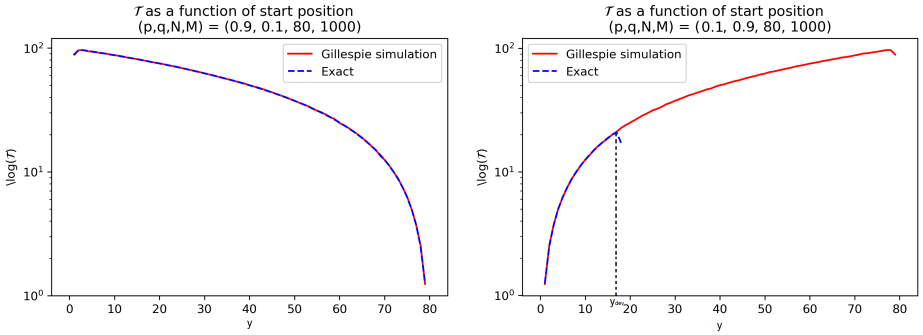


Figure 8: The exact results of eqs. (4.2), (4.9) and (4.22) and the Gillespie simulations in one plot. The value for y is varied, which is plotted on the x-axis. The plots are made for two different combinations of constant p, q: The left column has values $p = 0.3, q = 0.7$ and on the right $p = 0.4, q = 0.6$. For all situations the length of the state space was $N = 80$ and the Gillespie simulations were performed with $M = 10000$ walkers.



(a) (b)

Figure 9: The exact results of eq. (4.9) and the Gillespie simulation in one plot. The value for y was varied, which is plotted on the x-axis. The plots are made for constant rates $p = 0.9$ on the left and $p = 0.1$ on the right. $q = 1 - p$.
The exact results for fig. 9b is not displayed from a certain value of y. This is because the exact times are negative for these values of y due to a numerical error. Such values cannot be displayed in a logarithmic plot. However, the plot serves its aim by showing that the equation works well for $p = 0.9$, in contrast to $p = 0.1$.
For both situations the length of the state space is $N = 80$ and the Gillespie simulation is performed with $M = 1000$ walkers.

with c a constant given by

$$c = \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{1}{p} \left(\frac{q}{p}\right)^j}{\sum_{\xi=0}^{N-1} \left(\frac{q}{p}\right)^\xi}. \tag{5.2}$$

Note the indication of parts (a) and (b) in eq. (5.1). In figs. 9a and 9b it can be seen that for all $y \in [0, 80]$ the order of $\mathcal{T}(y+1) - \mathcal{T}(y)$ has an upper limit of 10, denoted as $\mathcal{O}(10)$. Therefore $(a) - (b) = \mathcal{O}(10)$ for all $y \in [0, 80]$.

Now the orders of (a) and (b) are considered individually, starting with $(q - p) > 0$. This implies that $\frac{q}{p} > 1$. Therefore $\left(\frac{q}{p}\right)^y$ blows up for large values of y, so (a) and (b) blow up as well. For example for $(p, q) = (0.1; 0.9)$, $y = 20$ and a state space of length 80 ($N = 80$) we find that $(a) = (b) = \mathcal{O}(10^{19})$. Remember that subtracting (a) and (b) from each other must give a result of maximum order ten. Therefore the values of (a) and (b) must be known with a precision of $\mathcal{O}(10)$. Such precision requires special software which was not used in generating these plots and this caused the numerical errors.

However if $(q - p) < 0$, then $\frac{q}{p} < 1$. Therefore $\left(\frac{q}{p}\right)^y \to 0$ for large y and (a) and (b) do not blow up. For $y = 20$, $N = 80$ and $(p, q) = (0.9; 0.1)$ one finds that $(a) = \mathcal{O}(10^{-18})$ and $(b) = \mathcal{O}(1)$. Now $(a) - (b) = \mathcal{O}(1)$ which is the desired order. It can be seen that in this case, the values need not be known with such precision, since (a) is negligible with respect to (b). Therefore there is no numerical error for $q < p$.

A similar reasoning can be used to explain the deviation in the results of

$\mathcal{T}_{\mathrm{ub}}$. Therefore derived expressions in eqs. (4.9) and (4.22) can be used for any value of y if $p > q$, however they should be used carefully for large values of y if $p < q$.

5.2 General expressions with fixed initial position

However as described in section 2.3, the only initial position of interest for the application to the CRISPR-Cas system is $y = 1$. Therefore the focus is the expressions for $\mathcal{P}_{\mathrm{clv}}(1), \mathcal{T}(1)$ and $\mathcal{T}_{\mathrm{ub}}(1)$ given in eqs. (4.3), (4.10) and (4.23). These reduced expressions are verified by Gillespie simulation. Again the rates are assumed to be constant throughout the state space which sum to one for all positions, i.e. $p_x = p$ and $q_x = q = 1 - p$. The simulated and exact values are displayed as a function of p in fig. 10.

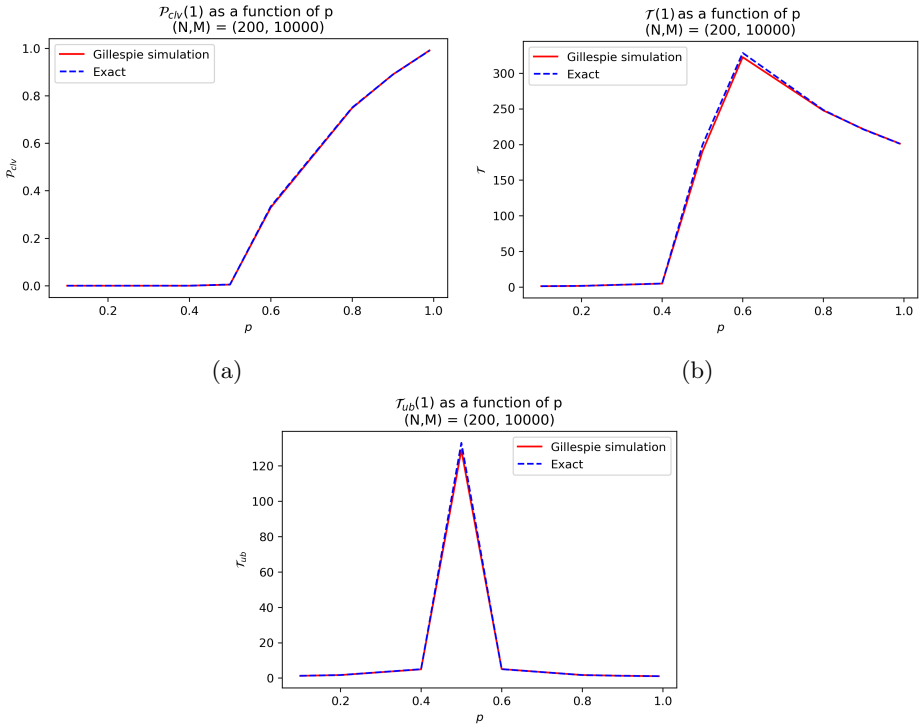


Figure 10: The exact results of eqs. (4.3), (4.10) and (4.23) and the Gillespie simulations in one plot. The plots are made for constant rates p and $q = 1 - p$. The value for p was varied, which is plotted on the x-axis. The initial position of the walkers was kept constant at 1.
Data points are $p = 0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9, 0.99$.
For all calculations the length of the state space was $N = 200$ and the Gillespie simulation was performed with $M = 10000$ walkers.

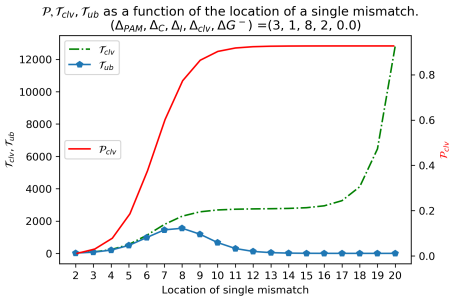
It can be seen that the results of the Gillespie simulations and the exact expressions overlap very well for all three quantities. According to the results of section 5.1, one could expect that for $q \gg p$ an error might appear. However

even for $(p, q) = (0.1; 0.9)$, the derived expressions give the expected results. This justifies the application of these expressions to the CRISPR-Cas system, in which the initial position is $y = 1$.

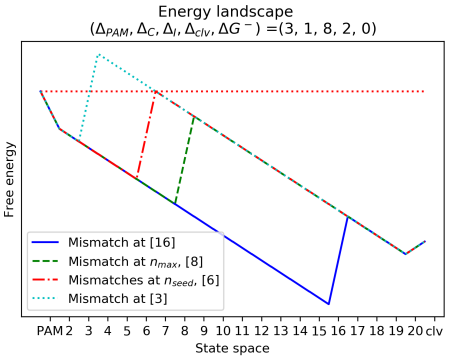
5.3 Single mismatch

In this section a state space with a single mismatch is considered. This mismatch is placed at varying locations in the state space and the effects on the unbinding and cleavage times are studied. The expression for $\mathcal{T}$ is left out of the analysis as it is not of practical use for the application to the CRISPR-Cas system.

Let us consider the state spaces with the energy landscapes as displayed in fig. 11b. The corresponding curves for $\mathcal{P}_{\text{clv}}(1), \mathcal{T}_{\text{ub}}(1), \mathcal{T}_{\text{clv}}(1)$ are displayed in fig. 11a.



(a) The values for $\mathcal{P}_{\text{clv}}(1), \mathcal{T}_{\text{clv}}(1), \mathcal{T}_{\text{ub}}(1)$ displayed as a function of the location of a single mismatch in the state space. The values for $\mathcal{P}_{\text{clv}}$ are displayed on the right vertical axis, the values for the times on the left vertical axis.
The maximum for $\mathcal{T}_{\text{ub}}$ is at position $n_{\text{max}} = 8$.



(b) Energy landscapes for the given parameters of fig. 11a with varying positions of the mismatch. The unbinding and cleavage times of the blue curve are relatively short. The cyan, dotted landscape has a short unbinding and long cleavage time. The green, dashed landscape has the maximum unbinding time. The red, dashed-dotted curve gives the energy landscape with a mismatch at n_{seed}. The dotted red line gives the initial free energy of the system.
n_{seed} was calculated with eq. (2.1), n_{max} was derived from fig. 11a.

Figure 11

Consider the cleavage time first. It can be seen that this time increases with an increasing location of the mismatch. This can be explained intuitively by referring to the energy landscape. If the mismatch is located close to the PAM, the Cas9 protein is not likely to pass the energy barrier of the mismatch since it is much higher than the barrier towards unbinding. However, if the protein passes this energy barrier, it will almost surely walk directly towards cleavage as the cleaved state is in an energy valley. Therefore the cleavage time is very low.

If the mismatch is located near the end of the R-loop, the energy landscape looks like the blue one in fig. 11b. In this landscape, the protein is very likely to cleave since the energy barrier to cleavage is much lower than the barrier to unbinding. The walker will, however, probably hop a long time in the energy valley just before the cleaved state before passing this barrier, as it is very high. Therefore the cleavage time grows with an increasing location of the mismatch.

Finally, if the mismatch is located at the last nucleotide, the barrier towards cleavage is extra high due to Δ_{clv}. Therefore it will take a longer time before Cas9 passed the single barrier and the cleavage time increases even further.

Now consider the unbinding time. First, if the mismatch is close to the PAM, the unbinding time is very low. This follows naturally from the cyan energy landscape in fig. 11b: the protein starts at the PAM and is very unlikely to pass the energy barrier. Therefore it will probably unbind. However it does not have a lot of space to move freely as the energy barrier is very close to the PAM. This implies that it cannot make a lot of steps before returning to the unbinding state. This causes a very low unbinding time.

Furthermore the unbinding time is very low if the mismatch is near the cleavage state. Let us consider the blue energy landscape in fig. 11b, which corresponds with a mismatch near the cleavage state. The walkers all start at the PAM, however the further the walkers move to the right, the less probable they are to return to the unbinding state. Therefore, the most probable way to arrive at the unbinding state is by starting at the PAM taking one step to the left to the unbound state immediately. Therefore it usually takes a short time to unbind if the mismatch is far from the PAM.

Now consider the maximum of the curve of the unbinding time in fig. 11a. According to the preceding paragraphs, the unbinding time increases by putting a mismatch far from the PAM such that the walkers have space to move, however it should not be too far away from the PAM else the walkers will not return to the unbound state. Therefore there must be a location which is a middle way of these two requirements, i.e. a mismatch at that location gives a longer unbinding time than a mismatch anywhere else on the state space. Let us define this location as $n_{\max}$. One can see in fig. 11b that $n_{\max} \neq n_{\text{seed}}$, as one might expect.

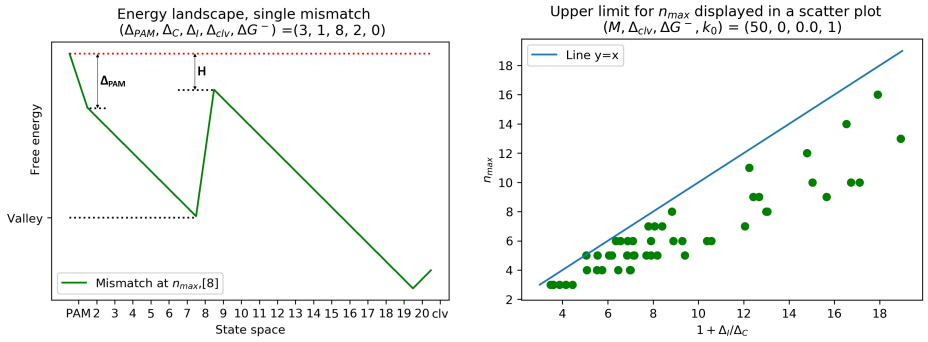
Consider the energy landscape in fig. 12a with a mismatch at $n_{\max}$. One can see that the energy barrier is higher than the energy level right after the PAM. H is defined as the difference between the initial energy of the system and the height of the energy barrier due to a mismatch at $n_{\max}$. Then H can be calculated by

$$H = \Delta_{\text{PAM}} + (n_{\max} - 1)\Delta_{\text{C}} - \Delta_{\text{I}}.$$

Notice $H < \Delta_{\text{PAM}}$ in fig. 12a. In that case the above expression can be rewritten to obtain

$$n_{\max} < 1 + \frac{\Delta_{\text{I}}}{\Delta_{\text{C}}}. \tag{5.3}$$

Figure 12b shows that this is not a coincidence; each dot in this scatter plot stands for a random combination of $\Delta_{\text{PAM}}, \Delta_{\text{C}}$ and Δ_{I}. For each combination, $n_{\max}$ was found by calculating the unbinding time for all possible positions of the mismatch and searching which location gave the maximum unbinding time. $n_{\max}$ was plotted versus the value of $1 + \frac{\Delta_{\text{I}}}{\Delta_{\text{C}}}$ and it clearly shows that this last value is an upper limit for $n_{\max}$.



(a) (b)

Figure 12: (a) The quantities H and Δ_{PAM} displayed in an energy landscape. The red, dotted line indicates the initial energy of the system. $n_{\max}$ was derived from fig. 11a
(b) The quantities of eq. (5.3) in a scatter plot. Each dot represents a random combination of $\Delta_{\mathrm{PAM}}, \Delta_{\mathrm{C}}, \Delta_{\mathrm{I}}$. Δ_{PAM} is a random integer between 1 and 9, $\Delta_{\mathrm{C}} \in [0.2; 1]$ and $\Delta_{\mathrm{I}} \in [2, 7]$. 1083 combination were drawn, hence 1083 dots are displayed.
The line $y = x$ is displayed, which represents the line at which $n_{\max}$ and $1 + \Delta_{\mathrm{I}}/\Delta_{\mathrm{C}}$ are equal. It can be seen that eq. (5.3) holds.
$\Delta_{\mathrm{clv}}, \Delta G^-, k_0$ are kept constant at $0, 0, 1$ respectively.

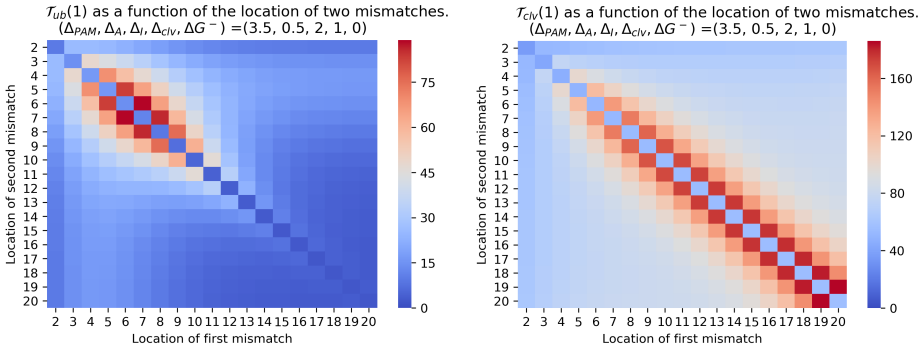
If $H < \Delta_{\mathrm{PAM}}$, the energy costs of reaching the PAM are lower than the costs of passing the barrier seen from the energy valley. Therefore a walker is more likely to reach the PAM instead of passing the mismatch. Apparently this is the optimum between giving the walkers enough space to walk and making sure enough walkers return to the unbinding state. This optimum results in a high expected unbinding time.

In general one can say that the cleavage time increases if the mismatch moves further away from the PAM. The unbinding time has a maximum for a mismatch at a certain location $n_{\max}$. This $n_{\max}$ is bounded from above by $1 + \frac{\Delta_{\mathrm{I}}}{\Delta_{\mathrm{C}}}$.

5.4 Double mismatch

Consider a state space with two mismatches. It is possible to display the unbinding time and cleavage time in a heatmap as a function of the locations of the two mismatches. They are displayed in fig. 13.

First the cleavage time is considered. It can be observed from fig. 13b that the closer two mismatches are, the longer it takes for a walker to cleave. This also follows from the energy landscape displayed in fig. 14. Looking at the green energy landscape in which the mismatches are placed at nearly subsequent locations, one can see that once a walker passed the first mismatch, it is very unlikely to directly pass the second mismatch. Instead the walker is very likely to fall back into the valley again and it has to start all over. Following this reasoning, it is expected to take a long time before the walker is able to pass both mismatches. In contrast, if two mismatches are placed far apart as displayed in the red energy landscape, the walker has some free space to move between two mismatches. Therefore it is less likely to fall back into the valley once it passed



(a) (b)

Figure 13: Heatmaps of $\mathcal{T}_{\mathrm{ub}}(1), \mathcal{T}_{\mathrm{clv}}(1)$ as a function of the location of two mismatches on the state space. If the two mismatches are at the same position, it is seen as one mismatch. That is why the diagonal has a relatively low unbinding and cleavage time.

the first mismatch, which reduces the cleavage time significantly. In the end, it is clear that a block of mismatches increases the cleavage time compared to spread mismatches.

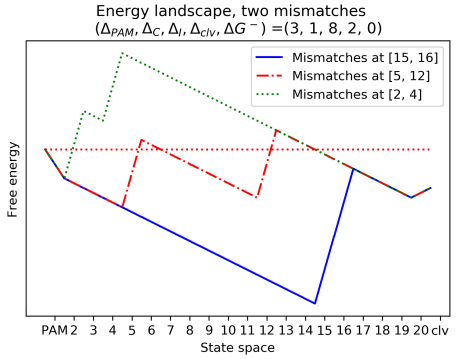


Figure 14: Energy landscapes with two mismatches. There is a block of two mismatches in the green, dotted and the blue landscapes. The red dashed curve has two spread mismatches.

Similarly, fig. 13a shows that the unbinding time increases significantly if two mismatches are close together. However, this only occurs when this block of mismatches is at the right distance from the PAM. This can be explained by the same reasoning as used for a single mismatch. If this block is too close to the PAM, the walkers do not have any space to move. Therefore the unbinding time is short. However, if the block is placed too far away, walkers are less likely to unbind. Therefore the only walkers that unbind are those that move from the PAM to the unbound state directly; otherwise it costs too much energy to reach the unbound state.

An upper limit for the location of a double mismatch for which $\mathcal{T}_{\mathrm{ub}}(1)$ has a maximum, can be found in the case of a double mismatch as well. Clearly, the

unbinding time is greatest in case of two subsequent mismatches so only these situations are considered. $n_{\max}$ is defined as the location of the last mismatch and H is defined as the energy level due to two subsequent mismatches as displayed in fig. 15a. H can be calculated by:

$$H = \Delta_{\mathrm{PAM}} + (n_{\max} - 1)\Delta_{\mathrm{C}} - 2\Delta_{\mathrm{I}}.$$

One can see in fig. 15b that $H < \Delta_{\mathrm{PAM}}$. Plugging this into the definition of H gives

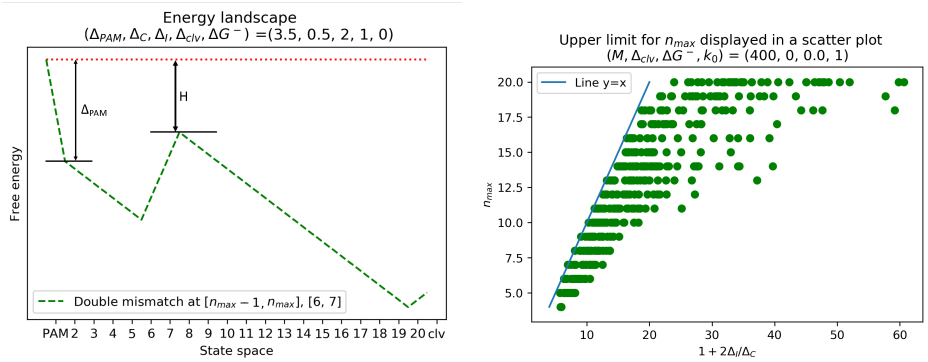
$$n_{\max} < 1 + \frac{2\Delta_{\mathrm{I}}}{\Delta_{\mathrm{C}}}. \tag{5.4}$$

This expression gives an upper limit for $n_{\max}$ with two mismatches. It can be seen in the scatter plot in fig. 15b that this upper limit holds for a random choice of parameters.

This result can be generalised to an energy landscape with a block of B subsequent mismatches. Then the following upper limit for $n_{\max}$ holds:

$$n_{\max} < 1 + \frac{B\Delta_{\mathrm{I}}}{\Delta_{\mathrm{C}}}. \tag{5.5}$$

It can be concluded that a block of mismatches increases the unbinding time only if the block is at the right distance from the PAM. A general expression for this upper limit, for any size of the block, is given by eq. (5.5).



(a) (b)

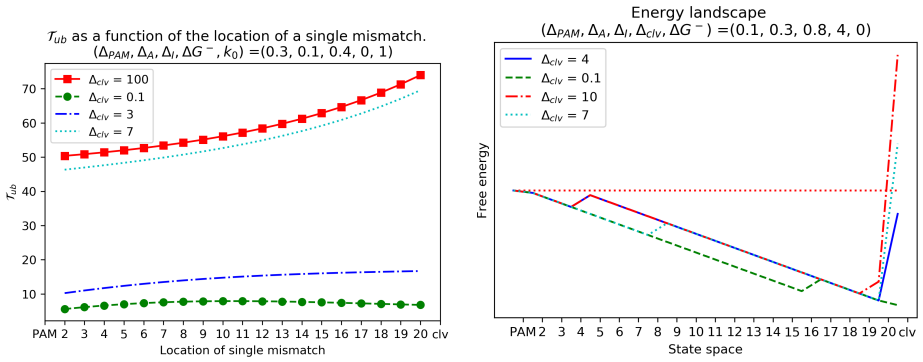
Figure 15: (a) The quantities H and Δ_{PAM} displayed in an energy landscape. The red, dotted line indicates the initial energy of the system. $n_{\max}$ was derived from fig. 13a
(b) The quantities of eq. (5.4) in a scatter plot. Each dot represents a random combination of $\Delta_{\mathrm{PAM}}, \Delta_{\mathrm{C}}, \Delta_{\mathrm{I}}$. Δ_{PAM} is a random integer between 1 and 9, $\Delta_{\mathrm{C}} \in [0.2; 1]$ and $\Delta_{\mathrm{I}} \in [2, 7]$. $n_{\max}$ was found by calculating the unbinding time for every position of two subsequent mismatches with these parameters and taking the position of the maximum. 400 combination were drawn, hence 400 dots are displayed.
The line $y = x$ is displayed, which represents the line at which $n_{\max}$ and $1 + 2\Delta_{\mathrm{I}}/\Delta_{\mathrm{C}}$ are equal. It can be seen that eq. (5.4) holds.
$\Delta_{\mathrm{clv}}, \Delta G^{-}, k_0$ are kept constant at $0, 0, 1$ respectively.

5.5 Influence of cleavage costs

Finally the influence of the parameter Δ_{clv} on the unbinding time is considered. The times are studied on a state space with a single mismatch of which its position is varied.

Figures 16a and 16b show that the unbinding time is significantly influenced by Δ_{clv} if the energy barrier towards cleavage approximates the value of the initial energy. One can see that the curves for $\Delta_{\text{clv}} = 0.1$ and $\Delta_{\text{clv}} = 3$ are very similar. These curves follow the shape as described in section 5.3. The curves for $\Delta_{\text{clv}} = 7$ and $\Delta_{\text{clv}} = 100$, however, differ significantly from the other two. For this combination of parameters, the energy barrier towards cleavage is higher than the energy barrier towards unbinding as one can observe in fig. 16b. Therefore walkers are more likely to unbind for any position of the mismatch. In that case the unbinding time still has a significant value for a mismatch close to cleavage because the walker is still likely to unbind.

It is remarkable that the first two curves are closer together than the latter two. This shows that the value of Δ_{clv} only influences the value of the unbinding time significantly if it causes the energy barrier to cleavage to be higher than the energy barrier to unbinding.



(a) (b)

Figure 16: Figure (a) shows $\mathcal{T}_{\text{ub}}(1)$ as a function of the location of a single mismatch for four different values of Δ_{clv} and constant values for the other parameters. Figure (b) shows the accompanying energy landscapes for the four different cases. It shows that the function is mainly influenced by whether the energy barrier to cleavage is lower or higher than the initial energy due to the value of Δ_{clv}.
The energy landscape for $\Delta_{\text{clv}} = 100$ is not shown due to the scaling. However, the landscape of $\Delta_{\text{clv}} = 10$ gives a good indication of the shape of the landscape.

6 Moment generating functions of stopping times

Up to now the main interest has been the expected stopping time of birth and death processes, i.e. the first moment. This first moment gives information about how long it usually takes a Cas9 protein to either cleave or unbind. However one could also be interested in higher moments of this stopping time, for example the variance and the kurtosis. All these moments are contained in the moment generating function which is often denoted as $\mathbb{E}[e^{\theta\tau}]$. This moment generating function provides full information on the distribution of the stopping time.

In this section two processes are considered:

1. A random walk with constant, equal rates: $p_x = q_x = p = 0.5$;
2. Brownian motion with drift.

For these cases the moment generating functions of the stopping time are found using martingales. First a short introduction to martingales is given. Next, the moment generating functions for the two cases are derived. Finally the full moment generating functions are verified using an inverse Laplace transform.

6.1 An introduction to martingales

There are two ways to classify random processes, being Markov chains and martingales [9]. Up to now all derivations have been done using the Markov property. This property states that the future of a Markov process is, given its present, independent of its past.

Martingales contrast this property as their future does depend on the past. Before introducing the definition of a martingale the following notation is introduced:

Definition 6.1. *Given the random process X_n and its outcomes on times $0, 1, \ldots, n$, denoted as $i_0, i_1, \ldots, i_n$ respectively. Then the set $\mathcal{F}_n$ is defined as follows:*

$$\mathcal{F}_n := \sigma\{X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\}$$

This set $\mathcal{F}_n$ is a short way to write down the sigma algebra of all past outcomes of the process X_n. Informally one could say that $\mathcal{F}_n$ describes the *information* contained in $X_0, \ldots, X_n$.

If some other random variable, let us say M_n, is dependent on this information $\mathcal{F}_n$, then M_n is called $\mathcal{F}_n$*-adapted* [9]. This means that for each $n \in \mathbb{N}$, there is some deterministic function g_n such that $M_n = g_n(X_0, \ldots, X_n)$.

Now the definition of a martingale is given [14]:

Definition 6.2 (Martingale)**.** *A random process M_n is called a martingale if:*

1. *M_n is $\mathcal{F}_n$-adapted;*
2. *$\forall n \in \mathbb{N}_0, \mathbb{E}[M_n] < \infty$;*
3. *$\forall n \in \mathbb{N}_0, \mathbb{E}[M_n | \mathcal{F}_{n-1}] = M_{n-1}$*

From the third property one could say that the expected outcome of a martingale's future is its present state. Let us introduce the tower property [2].

Lemma 6.3 (Tower property)**.** *Given a σ-algebra $\mathcal{G}$ and a random variable X, the following holds:*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$$

Using this tower property, an important property of a martingale can be derived:

Proposition 6.4. *Let M_n be a martingale, then the following holds:*

$$\mathbb{E}[M_n] = \mathbb{E}[M_0]$$

Proof. Assume M_n to be a martingale. Using the tower property (lemma 6.3) in step (1) and the definition of a martingale in (2) one can write:

$$\mathbb{E}[M_n] \overset{(1)}{=} \mathbb{E}\big[\mathbb{E}[M_n|\mathcal{F}_{n-1}]\big] \overset{(2)}{=} \mathbb{E}[M_{n-1}]$$

Iterating this result gives:

$$\mathbb{E}[M_n] = \mathbb{E}[M_{n-1}] = \ldots = \mathbb{E}[M_0]$$

□

This proposition 6.4 will play an important role in the remainder of this chapter.

Furthermore the following property of martingales can be shown:

Proposition 6.5. *Let M_n, N_n be two $\mathcal{F}_n$-adapted martingales, then $(M_n + N_n)$ is an $\mathcal{F}_n$-adapted martingale.*

Proof. The three defining properties of a martingale from definition 6.2 are proved:

1. Since M_n and N_n are $\mathcal{F}_n$-adapted, there exist some functions g_n, h_n such that $M_n = g_n(X_0, \ldots, X_n)$ and $N_n = h_n(X_0, \ldots, X_n)$ for all $n \in \mathbb{N}_0$. Define $f_n = g_n + h_n$. Then $M_n + N_n = (g_n + h_n)(X_0, \ldots, X_n) = f_n(X_0, \ldots, X_n)$. So indeed $(M_n + N_n)$ is $\mathcal{F}_n$-adapted.
2. By the definition of a martingale, $\mathbb{E}[M_n], \mathbb{E}[N_n] < \infty$ for all $n \in \mathbb{N}_0$. Since the expectation operator is linear one finds $\mathbb{E}[M_n + N_n] = \mathbb{E}[M_n] + \mathbb{E}[N_n] < \infty$.
3. By the linearity of the conditional expectation operator and since M_n and N_n are martingales, one can see that

$$\mathbb{E}[M_n + N_n|\mathcal{F}_{n-1}] = \mathbb{E}[M_n|\mathcal{F}_{n-1}] + \mathbb{E}[N_n|\mathcal{F}_{n-1}] = M_{n-1} + N_{n-1}$$

□

Finally the Dominated Convergence Theorem [5] is introduced, which will be used in the next sections to stop the martingales.

Theorem 6.6 (Dominated Convergence Theorem (DCT))**.** *Suppose that $\lim_{n\to\infty} X_n = X$ and there exists a random variable Y such that $\mathbb{E}[Y] < \infty$ and $|X_n| \leq Y$ for $n \geq 0$. Then*

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

6.2 Application to a random walk

This theory can be applied to a special case of the random walk to find the moment generating function of its stopping time. Assume a discrete state space $\{-a, -a+1, \ldots, -1, 0, 1, \ldots, b\}$ with a walker that starts at 0 (see fig. 17).

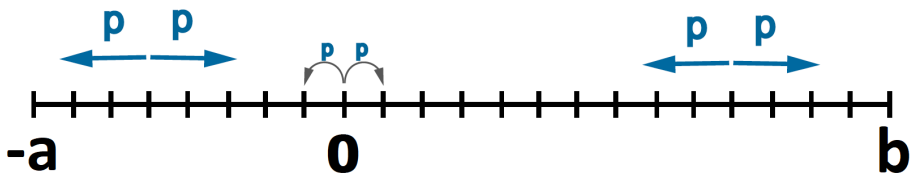


Figure 17: A scheme of the state space of this random walk. The walker starts at position 0 and has a probability $p = 0.5$ to move left and an equal probability to move right. The absorbers are positioned at $-a, b$.

At every time $n \in \mathbb{N}$ the walker has a probability of $p = 0.5$ to move left and an equal probability p to move right. This step is denoted as X_n, and this random variable can be summarised as follows:

$$X_n = \begin{cases} +1 & \text{with probability } p = 0.5 \\ -1 & \text{with probability } p = 0.5. \end{cases} \tag{6.1}$$

Furthermore S_n is defined. It gives the position of the walker at time n:

$$S_n = \sum_{i=0}^{n} X_i. \tag{6.2}$$

The random walk stops as soon as it reaches one of the absorbers at positions $-a$ and b. The length of time it took the walker to reach one of these absorbers is called the stopping time, which is denoted as

$$\tau := \inf\{n \in \mathbb{N} : S_n \in \{-a, b\}\} \tag{6.3}$$

Finally for any $\lambda, \alpha, \beta \in \mathbb{R}$ the following random variable is defined [8]:

$$\begin{aligned} M_n = &\exp\Big\{\lambda(S_n - \alpha) - n\log\big[\cosh\lambda\big]\Big\} \\ &+ \exp\Big\{-\lambda(S_n - \beta) - n\log\big[\cosh\lambda\big]\Big\} \end{aligned} \tag{6.4}$$

It can be shown that M_n is a martingale. A brief proof is given here, however the interested reader may want to read the detailed derivation in appendix B.1.

First one should notice that M_n is a deterministic function of the random variable S_n. Therefore M_n is $\mathcal{F}_n$-adapted with $\mathcal{F}_n = \sigma\{S_0, \ldots, S_n\}$. Let us define

$$N_{\pm,n} := \exp\Big\{\pm\lambda S_n - n\log\big[\cosh\lambda\big]\Big\}.$$

This random variable is $\mathcal{F}_n$-adapted as well. Notice that:

$$M_n = e^{-\lambda\alpha}N_{+,n} + e^{\lambda\beta}N_{-,n}. \tag{6.5}$$

Therefore, by proposition 6.5, M_n is a martingale if $N_{+,n}$ and $N_{-,n}$ are martingales. Using the definitions of X_n and S_n (eqs. (6.1) and (6.2)) it can be shown that

$$\begin{aligned}\mathbb{E}[N_{\pm,n}|\mathcal{F}_n] :&= \mathbb{E}\left[e^{\pm\lambda S_n - n\log[\cosh\lambda]}\middle|\mathcal{F}_n\right] \\ &= \left(pe^{\mp\lambda} + pe^{\pm\lambda}\right)e^{\pm\lambda S_{n-1} - n\log[\cosh\lambda]} \\ &= e^{\pm\lambda S_{n-1} - (n-1)\log[\cosh\lambda]} = N_{\pm,n-1}\end{aligned}$$

Therefore $N_{\pm,n}$ is a martingale, and so is M_n.

This martingale M_n will allow us to find the moment generating function of the stopping time for this special case of the random walk. This is done in three steps:

1) Proof that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$. Since the stopping time τ is not bounded, there is a probability that the walker moves on the state space for ever. In that case it cannot just be assumed that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$, even though M_n is a martingale. It is proved that this equality can be assumed by using the Dominated Convergence Theorem in theorem 6.6.

The following constant $k > 0$ and the notation $(a \wedge b) := \inf\{a, b\}$ are introduced. Then it can be shown by the DCT that the step marked by * in the following equation is allowed.

$$\mathbb{E}[M_\tau] \overset{(1)}{=} \mathbb{E}\left[\lim_{k\to\infty} M_{\tau\wedge k}\right] \overset{*}{=} \lim_{k\to\infty}\mathbb{E}[M_{\tau\wedge k}] \overset{(2)}{=} \lim_{k\to\infty}\mathbb{E}[M_0] = \mathbb{E}[M_0]. \tag{6.6}$$

Step (1) is allowed by the definition of the infimum. Step (2) is valid due to proposition 6.4 and since $(\tau \wedge k)$ is finite.

In order to use the DCT it must be shown that $|M_{\tau\wedge k}|$ is bounded. First of all it must be noticed that S_n is bounded. The walker is absorbed at positions $-a \leq 0$ or $b \geq 0$ and it starts at position 0, therefore it is clear that $S_n \in [-a, b] \cap \mathbb{Z}$ and since this set is a bounded set, S_n is bounded as well. Now the following can be defined for any given $\lambda, \alpha, \beta \in \mathbb{R}$:

$$C = \max_{S_n\in[-a,b]}\left\{e^{\lambda(S_n-\alpha)} + e^{-\lambda(S_n-\beta)}\right\} \geq 0 \tag{6.7}$$

Furthermore, since $p = 0.5$ and $0 < \frac{1}{\cosh(x)} \leq 1$ for all $x \in \mathbb{R}$, the following inequalities hold:

$$\begin{aligned}|M_{\tau\wedge k}| &= \left|\left[e^{\lambda(S_{\tau\wedge k}-\alpha)} + e^{-\lambda(S_{\tau\wedge k}-\beta)}\right]e^{-(\tau\wedge k)\log(\cosh\lambda)}\right| \\ &= \left|\left[e^{\lambda(S_{\tau\wedge k}-\alpha)} + e^{-\lambda(S_{\tau\wedge k}-\beta)}\right](\cosh\lambda)^{-(\tau\wedge k)}\right| \\ &\leq \left|\left[e^{\lambda(S_{\tau\wedge k}-\alpha)} + e^{-\lambda(S_{\tau\wedge k}-\beta)}\right]\cdot 1^{(\tau\wedge k)}\right| \\ &\leq C\end{aligned}$$

From this it is clear that $|M_{\tau\wedge k}|$ is bounded. Furthermore since C is deterministic, $\mathbb{E}[C] = C < \infty$. This implies that the DCT can be applied to $M_{\tau\wedge k}$ and that the step marked with * in eq. (6.6) is valid.

2) Calculation of $\mathbb{E}[M_0]$. The derivation of the moment generating function will make use of eq. (6.6). Therefore $\mathbb{E}[M_0]$ should be calculated.

First consider S_0. Since the walker starts at position 0 and the walker has not moved at time 0, it can be said that $S_0 = 0$. Since λ, α, β are deterministic, one finds

$$\mathbb{E}[M_0] = e^{-\lambda\alpha} + e^{\lambda\beta}. \tag{6.8}$$

3) Stopping the martingale. Combining eqs. (6.6) and (6.8) gives the following equation:

$$\mathbb{E}[M_\tau] = e^{-\lambda\alpha} + e^{\lambda\beta} \tag{6.9}$$

Solving this equation gives the desired moment generating function. Some steps of this derivation are displayed here, however the full derivation is given in appendix B.2.

First of all, by the definition of τ, the walker must have arrived at one of the two absorbers at the stopping time τ. This implies that $S_\tau = -a$ or $S_\tau = b$. Furthermore α, β are chosen such that $\alpha = \beta = \frac{b-a}{2}$. This information can be used to rewrite eq. (6.9) to

$$\mathbb{E}\left[e^{-\tau \log[\cosh \lambda]}\right] \cdot 2\cosh\left(\lambda\frac{a+b}{2}\right) = 2\cosh\left(\lambda\frac{b-a}{2}\right).$$

Since the desired expression is of the form $\mathbb{E}[e^{-\theta\tau}]$, θ is defined by $\theta = \log(\cosh\lambda)$. Then the desired moment generating function is given by:

$$\mathbb{E}\left[e^{-\theta\tau}\right] = \frac{\cosh\left(\frac{b-a}{2}\lambda(\theta)\right)}{\cosh\left(\frac{a+b}{2}\lambda(\theta)\right)} \tag{6.10}$$

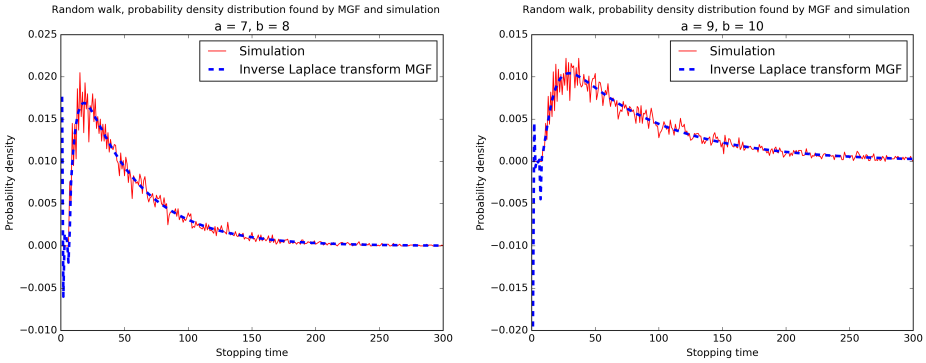
with

$$\lambda(\theta) = \theta + \log\left[1 + \sqrt{1 - e^{-2\theta}}\right]. \tag{6.11}$$

Validation by inverse Laplace transformation

Another way to derive the moment generating function of a random variable is by performing a Laplace transform to its probability density function (pdf) as these two are equivalent. This fact can be used to verify the derived MGF in this section. By applying an inverse Laplace transformation to the MGF, the pdf should be found. In fig. 18 the moment generating function was transformed numerically. Furthermore the random walk was simulated by Gillespie simulation.

It is clear that the two lines overlap very well in both situations. Therefore this moment generating function of the stopping time seems to be correct.



(a) (b)

Figure 18: Two plots verifying the moment generating function of the stopping time for different combinations of the parameters a and b. The inverse Laplace transform of the moment generating function is given by the blue dashed line, the simulation of the pdf by the red line. The inverse Laplace transform shows some oscillatory behaviour around 0 due to its complex nature.

6.3 Application to a Brownian motion with drift

Another limit case which is considered is a Brownian motion with drift. The theory of martingales can also be applied in this case to find the moment generating function of its stopping time.

A standard Brownian motion is defined as follows [10].

Definition 6.7. *A stochastic process W_t is called a standard Brownian motion if it has the following properties:*

1. $W_0 = 0$*;*
2. *The process has stationary increments, independent of time t;*
3. *For any time $0 \leq s < t$, $(W_t - W_s)$ is normally distributed with mean 0 and variance $(t - s)$.*

Consider the stochastic process $X_t = W_t + \mu t$, with W_t a standard Brownian motion. Then X_t is a Brownian motion with drif, the drift given by the coefficient $\mu \in \mathbb{R}$. X_t adheres to the first two properties described above. The third property changes as follows [13]:

3. For any time $0 \leq s < t$, $(X_t - X_s)$ is normally distributed with mean $\mu \cdot (t - s)$ and variance $t - s$.

In contrast to the random walk a Brownian motion is defined on a continuous state space $[-a, b]$ with absorber at $-a$ and b. The stopping time τ is defined as the first time that one of the absorbers is reached:

$$\tau := \inf \left\{ t : X_t \in \{-a, b\} \right\} \tag{6.12}$$

Of interested is the moment generating function of this stopping time, $\mathbb{E}\left[e^{\theta\tau}\right]$. In order to derive this function, the following random variable is defined for any

$\alpha, \lambda \in \mathbb{R}$ [8]:

$$M_t = \exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t - \mu X_t \right\} \sinh(\lambda X_t - \alpha) \tag{6.13}$$

This random variable is a martingale. A detailed proof of this statement is given in appendix C.1. This martingale allows us to find the moment generating function of the stopping time for a Brownian motion with drift using the same steps as in the previous subsection.

1) Proof that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$ Similar to the case of the random walk, the stopping time τ is not necessarily bounded. Therefore it cannot be assumed that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$. The following equation needs to be proved by the Dominated Convergence theorem (theorem 6.6):

$$\mathbb{E}[M_\tau] \overset{(1)}{=} \mathbb{E}\left[\lim_{k\to\infty} M_{\tau\wedge k}\right] \overset{*}{=} \lim_{k\to\infty} \mathbb{E}[M_{\tau\wedge k}] \overset{(2)}{=} \lim_{k\to\infty} \mathbb{E}[M_0] = \mathbb{E}[M_0]. \tag{6.14}$$

In this equation the step marked by (1) holds by the definition of the infimum. Step (2) is valid due to proposition 6.4 and since $(\tau \wedge k)$ is finite. The step marked with * needs to be proved with the DCT by showing that $|M_{\tau\wedge k}|$ is bounded.

First of all X_t is bounded by $-a < 0$ and $b > 0$. Furthermore, since $\lambda, \alpha \in \mathbb{R}$, the following can be defined:

$$C = \max_{X_t \in [-a,b]} \left\{ e^{-\mu X_t} \sinh(\lambda X_t - \alpha) \right\}. \tag{6.15}$$

Finally, since $\tau \wedge k \geq 0$ and by assuming $|\lambda| > |\mu|$, the following holds:

$$\begin{aligned} |M_{\tau\wedge k}| &= \left| e^{-\frac{1}{2}(\lambda^2-\mu^2)(\tau\wedge k) - \mu X_t} \sinh(\lambda X_t - \alpha) \right| \\ &\leq \left| e^{-\frac{1}{2}(\lambda^2-\mu^2)0} \cdot C \right| \\ &\leq |C| \end{aligned}$$

From this it follows that $|M_{\tau\wedge k}|$ is bounded if $|\lambda| > |\mu|$. Therefore the step marked with * in eq. (6.14) is allowed by the Dominated Convergence Theorem. Therefore The equation $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$ holds.

2) Calculation of $\mathbb{E}[M_0]$ In order to use proposition 6.4, the initial expectation of M_0 has to be found. According to the first property of Brownian motion $X_0 = 0$. Therefore

$$\mathbb{E}[M_0] = \mathbb{E}\left[e^0 \sinh(-\alpha)\right] = \sinh(-\alpha). \tag{6.16}$$

3) Stopping the martingale From eq. (6.14) it follows that the equation

$$\mathbb{E}[M_\tau] = \sinh(-\alpha) \tag{6.17}$$

is valid. Solving this gives the desired moment generating function. A full derivation is given in appendix C.2, however a few steps are given here.

By the definition of τ, the walker must be at the positions $X_\tau = -a$ or $X_\tau = b$ at time τ. Using this, and by choosing α smartly, the following equation is found from eq. (6.17):

$$\mathbb{E}\left[e^{-\frac{1}{2}(\lambda^2-\mu^2)t}\right] = \frac{e^{-\mu a}\sinh(-\alpha)}{\sinh(-\lambda a - \alpha)}.$$

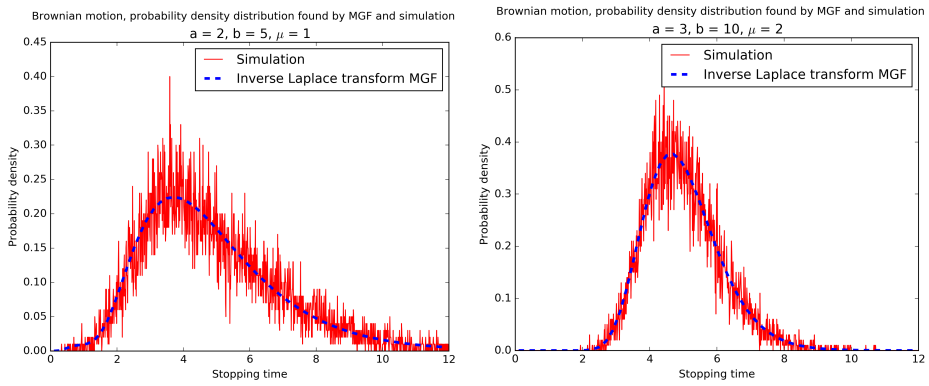
Since an expression for $\mathbb{E}[e^{-x\tau}]$ is to be found, the substitution $x = \frac{1}{2}(\lambda^2 - \mu^2)$ needs to be done. Recall from the assumptions of step 2) that $|\lambda| > |\mu|$, therefore $x > 0$. By substituting this expression and inserting the expression for α that was chosen, the desired moment generating function of the stopping time of a Brownian motion with drift is found:

$$\mathbb{E}[e^{-x\tau}] = \frac{e^{\mu b}\sinh(a\lambda(x)) + e^{-\mu a}\sinh(b\lambda(x))}{\sinh((a+b)\lambda(x))} \tag{6.18}$$

with $\lambda(x) = \sqrt{2x + \mu^2}$ in which $x > 0$.

Validation by inverse Laplace transformation

Analogous to the moment generating function of the random walk, the MGF of a Brownian motion with drift can be validated by an inverse Laplace transform. A simulated probability distribution function and the inverse Laplace transform of the derived MGF are plotted in fig. 19 for two combinations of a, b and μ.



(a) (b)

Figure 19: Two plots verifying the moment generating function of the stopping time for different combinations of a, b and μ. The inverse Laplace transform of the moment generating function is given by the blue dashed line, the simulation of the pdf by the red line.
The simulations were done with 10000 walkers and a time step of 0.01.

Again it can be seen that the inverse Laplace transform of the moment generating function fits very well through the simulated probability density functions. Therefore the MGF seems to be correct.

7 Conclusion

In this thesis closed form expressions for the cleavage probability and the cleavage and unbinding times were found for the CRISPR associated protein Cas9. Since such an R-loop is modelled by a birth and death process, the derivations could be done using Markov theory and semigroups.

Initially these expressions were derived for general birth and death processes. The expression for the probability was consistent with the simulations, however the expressions for the stopping times showed great deviations if the rates were chosen such that $q > p$. It was shown that these deviations were caused by a numerical error.

The derived expressions were reduced to a form useful for the application to the CRISPR-Cas system. They were analysed as a function of the location of one mismatch and it was shown that the unbinding time reaches a maximum. Moreover, it was noted that the cleavage time increases if the mismatch is positioned closer the cleaved state. Using the language of free-energy landscapes, these findings were rationalised and an upper limit for the position at which the unbinding time has a maximum was found.

Furthermore the expressions were analysed as a function of the location of two mismatches. It was shown that a block of two mismatches causes an increase of the cleavage time, compared to two mismatches spread throughout the state space. Similar to the observations in the analysis with one mismatch, the unbinding time reaches a maximum if the block of mismatches is placed at the right distance from the PAM. Again an upper limit for this distance was found for a general number of subsequent mismatches.

It was shown that the unbinding time as a function of the energy costs to cleave has two possible behaviours, depending on whether the energy barrier towards unbinding is higher or lower than the initial energy of the system. Any variations of these energy costs within the two possibilities did not influence the resulting unbinding time significantly.

Finally the moment generating functions of the stopping time were derived for two special Markov processes; a random walk and a Brownian motion with drift. These were derived using martingales. They were validated by applying a numerical inverse Laplace transform to them and comparing them with a simulated probability density function. This comparison showed that the method results in correct moment generating functions.

8 Outlook

This thesis contains closed form expressions for the cleavage and unbinding times of a CRISPR associated protein. The temporal information given by these expressions could be very useful in lab experiments and medical applications. It might take a long time before the protein cleaves or unbinds and a result can be seen in the experiment. Therefore one might want to know what time it takes before cleavage or unbinding is done and set up his lab experiment accordingly.

Furthermore the expressions could be very useful to test the validity of the model presented by Klein et al. In their paper, the expressions for the cleavage probability have been fitted to the data from lab experiments. These fits matched the data and therefore it provided a validity check of the model. Fitting the derived equations to times obtained from the lab serves as a second validity check of the model. If the equations fit the data, it gives a double confirmation of the assumed model.

The expressions could also be used to derive a more generalised model, compared to the minimal model presented in this thesis. The parameters of the minimal model are merely based on whether the base pair is a match or a mismatch. The parameters of a generalised model are also position dependent and lab experiments show that this is indeed the case. Mismatches at certain positions have a smaller effect on the cleavage probability than other. This is not coherent with the minimal model. Since the expressions are also derived for a general set of $\Delta(j)$, they could be fitted to a dataset to study the position dependency of the influence of mismatches.

Furthermore it might be interesting to find the full distribution of the cleavage and unbinding times. This could be done by finding the moment generating function of the stopping time of a general birth and death process. This provides multiple layers to test the model and it gives the variance that might be useful for lab experiments and medical applications. To that end, a suitable martingale needs to be found.

A Solving the difference equations

A.1 Probability to cleave

In this section a closed expression for the probability to cleave given the initial position y. Consider a discrete state space with an absorbers at $x = 0$ and $x = N$. Recall the definitions of $\mathcal{P}_{\text{clv}}$ and the generator L from section 4 and eq. (3.10):

$$\mathcal{P}_{\text{clv}}(y) := \mathbb{P}_y(T_N < T_0) \tag{A.1}$$

$$Lf(x) = p_x(f(x+1) - f(x)) - q_x(f(x) - f(x-1)) \tag{A.2}$$

From the definition of $\mathcal{P}_{\text{clv}}$ it is clear that the boundary conditions are $\mathcal{P}_{\text{clv}}(0) = 0$ and $\mathcal{P}_{\text{clv}}(N) = 1$; after all when starting at $x = N$ the time to reach $x = N$ is surely smaller than the time to reach $x = 0$. In section 4.1 it was derived that the difference equation $L\mathcal{P}_{\text{clv}}(x) = 0$ holds. In this appendix this difference equation is solved.

The first step is to plug in the definition of L into the difference equation:

$$\mathcal{P}_{\text{clv}}(x+1) - \mathcal{P}_{\text{clv}}(x) = \frac{q(x)}{p(x)} \left[\mathcal{P}_{\text{clv}}(x) - \mathcal{P}_{\text{clv}}(x-1)\right].$$

Plugging this result in itself one finds

$$\mathcal{P}_{\text{clv}}(x+1) - \mathcal{P}_{\text{clv}}(x) = \frac{q(x)}{p(x)} \frac{q(x-1)}{p(x-1)} \left[\mathcal{P}_{\text{clv}}(x-1) - \mathcal{P}_{\text{clv}}(x-2)\right]$$

which, by iterating, eventually gives

$$\begin{aligned} \mathcal{P}_{\text{clv}}(x+1) - \mathcal{P}_{\text{clv}}(x) &= \frac{q(x)}{p(x)} \frac{q(x-1)}{p(x-1)} \cdots \frac{q(1)}{p(1)} \left[\mathcal{P}_{\text{clv}}(1) - \mathcal{P}_{\text{clv}}(0)\right] \\ &= \prod_{\eta=1}^{x} \left(\frac{q_\eta}{p_\eta}\right) \cdot \left[\mathcal{P}_{\text{clv}}(1) - \mathcal{P}_{\text{clv}}(0)\right] \end{aligned} \tag{A.3}$$

Assuming y is the walker's starting point, we are interested in $\mathcal{P}_{\text{clv}}(y)$. Therefore eq. (A.3) is summed over x from 0 to $y-1$ to obtain $\mathcal{P}_{\text{clv}}(y)$ in the equation. This results in a telescope series on the left hand side:

$$\sum_{x=0}^{y-1} \left(\mathcal{P}_{\text{clv}}(x+1) - \mathcal{P}_{\text{clv}}(x)\right) = \sum_{x=0}^{y-1} \left(\prod_{\eta=1}^{x} \frac{q_\eta}{p_\eta}\right) \cdot \left[\mathcal{P}_{\text{clv}}(1) - \mathcal{P}_{\text{clv}}(0)\right] \tag{A.4}$$

$$\mathcal{P}_{\text{clv}}(y) - \mathcal{P}_{\text{clv}}(0) = \sum_{x=0}^{y-1} \left(\prod_{\eta=1}^{x} \frac{q_\eta}{p_\eta}\right) \cdot \left[\mathcal{P}_{\text{clv}}(1) - \mathcal{P}_{\text{clv}}(0)\right] \tag{A.5}$$

Using the boundary conditions $\mathcal{P}_{\text{clv}}(0) = 0$ and $\mathcal{P}_{\text{clv}}(N) = 1$, an expression for $\mathcal{P}_{\text{clv}}(1)$ can be found. By plugging this expression into eq. (A.5) it follows that

$$\mathcal{P}_{\text{clv}}(y) = \frac{\sum_{x=0}^{y-1} \left(\prod_{\eta=1}^{x} \frac{q_\eta}{p_\eta}\right)}{\sum_{x=0}^{N-1} \left(\prod_{\eta=1}^{x} \frac{q_\eta}{p_\eta}\right)}$$

Using the definition of $\varphi(x)$ given in definition 4.3, this expression can be reduced to

$$\mathcal{P}_{\text{clv}}(y) = \frac{\sum_{x=0}^{y-1} \varphi(x)}{\sum_{x=0}^{N-1} \varphi(x)}. \tag{A.6}$$

Hence a closed expression for $\mathcal{P}_{\text{clv}}(y)$ was found.

A.2 Expected time to cleave or unbind

Consider a birth and death process on a discrete state space with absorbers at $x = 0$ and $x = N$. A closed expression for the expected time it takes a walker to reach one of its absorbers, given its initial position y is derived. Recall the definition of $\mathcal{T}$ from section 4:

$$\mathcal{T}(y) = \mathbb{E}(T_{0,N}) \tag{A.7}$$

It follows naturally that the boundary conditions are $\mathcal{T}(0) = \mathcal{T}(N) = 0$: if the walker starts at $x = 0$ or N, the time to reach $x = 0$ or $x = N$ is zero. In section 4.2 it was derived that the difference equation $L\mathcal{T} = -1$ holds. In this appendix this difference equation is solved.

Using definition of the generator L given in eq. (A.2), it can be seen that:

$$\mathcal{T}(x+1) - \mathcal{T}(x) = \frac{q_x}{p_x}[\mathcal{T}(x) - \mathcal{T}(x-1)] - \frac{1}{p_x} \tag{A.8}$$

Notice that this equation can be plugged in itself as was done in the previous section. Writing out two iterations gives:

$$\begin{aligned}
\mathcal{T}(x+1) - \mathcal{T}(x) &= \frac{q_x}{p_x}\left(\frac{q_{x-1}}{p_{x-1}}\left[\mathcal{T}(x-1) - \mathcal{T}(x-2)\right] - \frac{1}{p_{x-1}}\right) - \frac{1}{p_x} \\
&= \frac{q_x q_{x-1}}{p_x p_{x-1}}\left[\mathcal{T}(x-1) - \mathcal{T}(x-2)\right] - \frac{q_x}{p_x p_{x-1}} - \frac{1}{p_x} \\
&\vdots \\
&= \frac{q_x q_{x-1} q_{x-2}}{p_x p_{x-1} p_{x-2}}\left[\mathcal{T}(x-2) - \mathcal{T}(x-3)\right] - \frac{q_x q_{x-1}}{p_x p_{x-1} p_{x-2}} - \frac{q_x}{p_x p_{x-1}} - \frac{1}{p_x}
\end{aligned}$$

Generalizing the result and using definition 4.3, one finds:

$$\mathcal{T}(x+1) - \mathcal{T}(x) = \varphi(x)[\mathcal{T}(1) - \mathcal{T}(0)] - \sum_{i=0}^{x-1}\left\{\frac{1}{p_{x-i}}\frac{\varphi(x)}{\varphi(x-i)}\right\} \tag{A.9}$$

Since an expression for $\mathcal{T}(y)$ is of interest, eq. (A.9) is summed over x from 0 to $y-1$. This gives a telescope series on the left hand side and one finds:

$$\mathcal{T}(y) - \mathcal{T}(0) = \sum_{x=0}^{y-1}\left\{\varphi(x)\left[\mathcal{T}(1) - \mathcal{T}(0)\right] - \sum_{i=0}^{x-1}\left(\frac{1}{p_{x-i}}\frac{\varphi(x)}{\varphi(x-i)}\right)\right\} \tag{A.10}$$

Using the boundary conditions $\mathcal{T}(0) = \mathcal{T}(N) = 0$, eq. (A.10) can be rewritten to find a closed expression for $\mathcal{T}(y)$:

$$\mathcal{T}(y) = \sum_{x=0}^{y-1}\left\{\frac{\sum_{\xi=0}^{N-1}\sum_{j=0}^{\xi-1}\frac{1}{p_{\xi-j}}\frac{\varphi(\xi)}{\varphi(\xi-j)}}{\sum_{\xi=0}^{N-1}\varphi(\xi)}\varphi(x) - \sum_{i=0}^{x-1}\frac{1}{p_{x-i}}\frac{\varphi(x)}{\varphi(x-i)}\right\}. \tag{A.11}$$

A.3 Expected time to cleave

A closed expression for the time to cleave, given the initial position y can be derived. Consider a discrete state space with absorbers at $x = 0$ and $x = N$. Recall the definition of $\mathcal{T}_{\text{clv}}(y)$ from section 4:

$$\mathcal{T}_{\text{clv}}(y) = \mathbb{E}_y(T_{0,N}|T_N < T_0) \tag{A.12}$$

To derive an expression for $\mathcal{T}_{\text{clv}}(y)$, the following identity is used:

$$\mathbb{E}_y(T_{0,N}|T_N < T_0) = \frac{\mathbb{E}_y\left(T_{0,N} I(T_N < T_0)\right)}{\mathbb{P}_y(T_N < T_0)}.$$

in which $I(.)$ is the indicator function. Let us define $\kappa_y := \mathbb{E}_y(T_{0,N} I(T_N < T_0))$, this identity can be rewritten as

$$\mathcal{T}_{\text{clv}}(y) = \frac{\kappa_y}{\mathcal{P}_{\text{clv}}(y)} \tag{A.13}$$

An expression for $\mathcal{P}_{\text{clv}}(y)$ was derived in appendix A.1, therefore an expression for $\kappa(y)$ needs to be found only. The boundary conditions for κ follow from its definition: they are $\kappa_0 = \kappa_N = 0$ as $T_{0,N} = 0$ for these two positions.

In section 4.3 the difference equation $L\kappa_x = -\mathcal{P}_{\text{clv}}(x)$ was derived. This difference equation is solved in this appendix. First use the definition of L given in eq. (A.2):

$$\kappa_{x+1} - \kappa_x = \frac{q_x}{p_x}[\kappa_x - \kappa_{x-1}] - \frac{1}{p_x}\mathcal{P}_{\text{clv}}(x)$$

Analogous to the previous appendices, this equation can be plugged into itself. Two iterations are worked out:

$$\begin{aligned}
\kappa_{x+1} - \kappa_x &= \frac{q_x}{p_x}\left[\frac{q_{x-1}}{p_{x-1}}[\kappa_{x-1} - \kappa_{x-2}] - \frac{1}{p_{x-1}}\mathcal{P}_{\text{clv}}(x-1)\right] - \frac{1}{p_x}\mathcal{P}_{\text{clv}}(x) \\
&= \frac{q_x q_{x-1}}{p_x p_{x-1}}[\kappa_{x-1} - \kappa_{x-2}] - \frac{q_x}{p_x p_{x-1}}\mathcal{P}_{\text{clv}}(x-1) - \frac{1}{p_x}\mathcal{P}_{\text{clv}}(x) \\
&= \frac{q_x q_{x-1} q_{x-2}}{p_x p_{x-1} p_{x-2}}[\kappa_{x-2} - \kappa_{x-3}] - \frac{q_x q_{x-1}}{p_x p_{x-1} p_{x-2}}\mathcal{P}_{\text{clv}}(x-2) - \frac{q_x}{p_x p_{x-1}}\mathcal{P}_{\text{clv}}(x-1) - \frac{1}{p_x}\mathcal{P}_{\text{clv}}(x)
\end{aligned}$$

Iterating this, the following general result is found:

$$\kappa_{x+1} - \kappa_x = \varphi(x)[\kappa_1 - \kappa_0] - \sum_{i=0}^{x-1} \frac{\varphi(x)}{\varphi(x-i)} \frac{1}{p_{x-i}} \mathcal{P}_{\text{clv}}(x-i) \tag{A.14}$$

Since we are interested in κ_y, eq. (A.14) should be summed over x from 0 to $y-1$. This results in a telescope series on the left hand side.

$$\kappa_y - \kappa_0 = \sum_{x=0}^{y-1} \left\{ \varphi(x)[\kappa_1 - \kappa_0] - \sum_{i=0}^{x-1} \frac{\varphi(x)}{\varphi(x-i)} \frac{1}{p_{x-i}} \mathcal{P}_{\text{clv}}(x-i) \right\}.$$

Using the boundary conditions $\kappa_0 = \kappa_N = 0$, the equation can be rewritten into a closed expression for κ_y:

$$\kappa_y = \sum_{x=0}^{y-1} \left\{ \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\text{clv}}(\xi-j)}{\sum_{\xi=0}^{N-1} \varphi(\xi)} \varphi(x) - \sum_{i=0}^{x-1} \frac{\varphi(x)}{\varphi(x-i)} \frac{1}{p_{x-i}} \mathcal{P}_{\text{clv}}(x-i) \right\}. \tag{A.15}$$

Then from eq. (A.13) one finds that the expression for the expected cleavage time is

$$\mathcal{T}_{\mathrm{clv}}(y) = \frac{1}{\mathcal{P}_{\mathrm{clv}}(y)} \sum_{x=0}^{y-1} \left\{ \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\mathrm{clv}}(\xi-j)}{\sum_{\xi=0}^{N-1} \varphi(\xi)} \varphi(x) - \sum_{i=0}^{x-1} \frac{\varphi(x)}{\varphi(x-i)} \frac{1}{p_{x-i}} \mathcal{P}_{\mathrm{clv}}(x-i) \right\}. \tag{A.16}$$

A.4 Expected time to unbind

The expected unbinding time $\mathcal{T}_{\mathrm{ub}}$ can be found by a similar approach as used for $\mathcal{T}_{\mathrm{clv}}$. Recall the definition for $\mathcal{T}_{\mathrm{ub}}(y)$ from section 4:

$$\mathcal{T}_{\mathrm{ub}}(y) = \mathbb{E}_y(T_{0,N} | T_0 < T_N). \tag{A.17}$$

Then, similar to eq. (A.13) the following identity holds for $\mathcal{T}_{\mathrm{ub}}(y)$:

$$\mathcal{T}_{\mathrm{ub}}(y) = \frac{\nu_y}{\mathcal{P}_{\mathrm{ub}}(y)} \tag{A.18}$$

with $\nu_y := \mathbb{E}_y(T_{0,N} I(T_0 < T_N))$. It is known that $\mathcal{P}_{\mathrm{ub}}(y) = 1 - \mathcal{P}_{\mathrm{clv}}(y)$ since, in the end, a walker will reach one of the absorbers. Therefore an expression for ν_y is to be found only. In section 4.4 the following difference equation was derived for ν:

$$L\nu_x = -\mathcal{P}_{\mathrm{ub}}(x).$$

This difference equation is similar to the equation for κ_x and therefore gives a similar result. It can be deduced that a closed expression for ν_y is:

$$\nu_y = \sum_{x=0}^{y-1} \left\{ \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\mathrm{ub}}(\xi-j)}{\sum_{\xi=0}^{N-1} \varphi(\xi)} \varphi(x) - \sum_{i=0}^{x-1} \frac{\varphi(x)}{\varphi(x-i)} \frac{1}{p_{x-i}} \mathcal{P}_{\mathrm{ub}}(x-i) \right\}.$$

Then from eq. (A.18) it follows that

$$\mathcal{T}_{\mathrm{ub}}(y) = \frac{1}{\mathcal{P}_{\mathrm{ub}}(y)} \sum_{x=0}^{y-1} \left\{ \frac{\sum_{\xi=0}^{N-1} \sum_{j=0}^{\xi-1} \frac{\varphi(\xi)}{\varphi(\xi-j)} \frac{1}{p_{\xi-j}} \mathcal{P}_{\mathrm{ub}}(\xi-j)}{\sum_{\xi=0}^{N-1} \varphi(\xi)} \varphi(x) - \sum_{i=0}^{x-1} \frac{\varphi(x)}{\varphi(x-i)} \frac{1}{p_{x-i}} \mathcal{P}_{\mathrm{ub}}(x-i) \right\} \tag{A.19}$$

with y the initial position of the walker.

B The moment generating function for the random walk

In section 6.2 the moment generating function of the stopping time of a special case of the random walk is derived. Two proofs are given in detail here. First it is proved that the used martingale is indeed a martingale. Secondly the derivation of the moment generating function is given.

B.1 Proof that M_n is a martingale

Recall the definition of M_n and $N_{\pm,n}$:

$$\begin{aligned} M_n :=& \exp\Big\{\lambda(S_n-\alpha) - n\log\big[\cosh\lambda\big]\Big\} \\ &+ \exp\Big\{-\lambda(S_n-\beta) - n\log\big[\cosh\lambda\big]\Big\} \\ N_{\pm,n} :=& \exp\Big\{\pm\lambda S_n - n\log\big[\cosh\lambda\big]\Big\} \end{aligned}$$

From these definitions it can be seen that $M_n = e^{-\lambda\alpha}N_{+,n} + e^{\lambda\beta}N_{-,n}$. Since both M_n and $N_{\pm,n}$ are deterministic functions dependent on the random variable S_n, it can be said that they are all $\mathcal{F}_n$-adapted with $\mathcal{F}_n = \sigma\{S_0,\ldots,S_n\}$. Therefore, using proposition 6.5, M_n is a martingale if $N_{+,n}$ and $N_{-,n}$ are martingales. To show that these are martingales, the three defining properties of a martingale are verified.

First, it was already explained that $N_{\pm,n}$ is $\mathcal{F}_n$-adapted. Second $\mathbb{E}[S_n]$ is finite since the state space is finite. Therefore, since α,β and λ are finite constants, it is clear that $\mathbb{E}[N_{\pm,n}] < \infty$.

Finally it must be shown that $\mathbb{E}[N_{\pm,n}|\mathcal{F}_n] = N_{\pm,n-1}$. One can see that

$$\mathbb{E}[N_{\pm,n}|\mathcal{F}_n] = \mathbb{E}\Big[\exp\big\{\pm\lambda S_n - n\log\left[\cosh\lambda\right]\big\}\big|\mathcal{F}_n\Big].$$

Using the definition of S_n in eq. (6.2) one finds

$$= \mathbb{E}\Big[\exp\big\{\pm\lambda(S_{n-1}-X_n) - n\log\big[\cosh\lambda\big]\big\}\big|\mathcal{F}_n\Big].$$

Since S_{n-1} and X_n are independent, the expectation values can be split. Furthermore, as $S_{n-1}\in\mathcal{F}_n$, the second expectation value is deterministic:

$$= \mathbb{E}\Big[\exp\big\{\mp\lambda X_n\big\}|\mathcal{F}_n\Big]\exp\big\{\pm\lambda S_{n-1} - n\log\left[\cosh\lambda\right]\big\}$$

X_n has outcomes ±1 with probability p (see eq. (6.1)). Therefore using the law of the subconscious statiscian [4], one can write this as:

$$\begin{aligned} &= \big(pe^{\mp}\lambda + pe^{\pm}\lambda\big)\cdot\exp\left\{\pm\lambda S_{n-1} - n\log\left[\cosh\lambda\right]\right\} \\ &= \cosh(\lambda)\cdot\exp\left\{\pm\lambda S_{n-1} - n\log\left[\cosh\lambda\right]\right\} \end{aligned}$$

Using several calculus rules one finds:

$$\begin{aligned}
&= \exp\{\log[\cosh\lambda]\} \cdot \exp\{\pm\lambda S_{n-1} - n\log[\cosh\lambda]\} \\
&= \exp\{\pm\lambda S_{n-1} - (n-1)\log[\cosh\lambda]\} \\
&= N_{\pm,n-1}
\end{aligned}$$

Therefore $N_{\pm,n}$ is a martingale. From this it follows that M_n is a martingale.

B.2 Derivation of the moment generating function

Consider a random walk on a discrete state space $\mathbb{S} = \{-a, -a+1, \ldots, 0, \ldots, b\}$. At each position the walker has an equal probability $p = 0.5$ to move left or right. In this appendix the moment generating function of the stopping time τ is derived for this situation.

Recall the martingale defined in eq. (6.4) as a function of S_n:

$$\begin{aligned}
M_n(S_n) = &\exp\Big\{\lambda(S_n - \alpha) - n\log\big[\cosh\lambda\big]\Big\} \\
&+ \exp\Big\{-\lambda(S_n - \beta) - n\log\big[\cosh\lambda\big]\Big\}.
\end{aligned}$$

By stopping the martingales in section 6.2 the following result was found (see eq. (6.9)):

$$\mathbb{E}[M_\tau] = e^{-\lambda\alpha} + e^{\lambda\beta}. \tag{B.1}$$

At the stopping time τ, the walker is, by the definition of τ, either at position $-a$ or b. Therefore $S_\tau \in \{-a, b\}$. $\mathbb{E}[M_\tau]$ can be written as follows:

$$\begin{aligned}
\mathbb{E}[M_\tau] &= \sum_{x\in\{-a,b\}} \mathbb{E}[M_\tau(S_\tau = x) \cdot I(S_\tau = x)] \\
&= \mathbb{E}\Big[\exp\Big\{\lambda(-a-\alpha) - \tau\log\big[\cosh\lambda\big]\Big\} I(S_\tau = -a)\Big] \\
&\quad + \mathbb{E}\Big[\exp\Big\{-\lambda(-a-\beta) - \tau\log\big[\cosh\lambda\big]\Big\} I(S_\tau = -a)\Big] \\
&\quad + \mathbb{E}\Big[\exp\Big\{\lambda(b-\alpha) - \tau\log\big[\cosh\lambda\big]\Big\} I(S_\tau = b)\Big] \\
&\quad + \mathbb{E}\Big[\exp\Big\{-\lambda(b-\beta) - \tau\log\big[\cosh\lambda\big]\Big\} I(S_\tau = b)\Big]
\end{aligned} \tag{B.2}$$

$\alpha, \beta \in \mathbb{R}$ are constants and they can be chosen freely. The following system of equations is solved for α, β:

$$\begin{cases} -a - \alpha = -(b - \beta) \\ -(-a - \beta) = b - \alpha \end{cases} \implies \alpha = \beta = \frac{-a+b}{2}$$

α, β are chosen as expressed above. Then one finds the following four expressions:

$$\begin{aligned}
-a - \alpha &= -\frac{a+b}{2} & \qquad b - \alpha &= \frac{a+b}{2} \\
-(b - \beta) &= -\frac{a+b}{2} & \qquad -(-a - \beta) &= \frac{a+b}{2}
\end{aligned}$$

Using these four expressions and plugging them into eq. (B.2), one finds:

$$\begin{aligned}\mathbb{E}[M_\tau] = \mathbb{E}\Bigg[\exp\Bigg\{ -\lambda\frac{a+b}{2} - \tau \log\big[\cosh\lambda\big]\Bigg\} I(S_\tau = -a)\Bigg] \\ + \mathbb{E}\Bigg[\exp\Bigg\{ \lambda\frac{a+b}{2} - \tau \log\big[\cosh\lambda\big]\Bigg\} I(S_\tau = -a)\Bigg] \\ + \mathbb{E}\Bigg[\exp\Bigg\{ \lambda\frac{a+b}{2} - \tau \log\big[\cosh\lambda\big]\Bigg\} I(S_\tau = b)\Bigg] \\ + \mathbb{E}\Bigg[\exp\Bigg\{ -\lambda\frac{a+b}{2} - \tau \log\big[\cosh\lambda\big]\Bigg\} I(S_\tau = b)\Bigg]. \end{aligned} \tag{B.3}$$

Since there are no other possibilities than $S_\tau = -a$ or $S_\tau = b$, it can be seen that $I(S_\tau = -a) + I(S_\tau = b) = 1$. Using this, eq. (B.3) can be simplified to

$$\begin{aligned}\mathbb{E}[M_\tau] &= \mathbb{E}\Big[e^{-\lambda\frac{a+b}{2} - \tau\log[\cosh\lambda]} + e^{\lambda\frac{a+b}{2} - \tau\log[\cosh\lambda]}\Big] \\ &= \mathbb{E}\Big[e^{-\tau\log[\cosh\lambda]}\Big] \cdot 2\cosh\left(\lambda\frac{a+b}{2}\right). \end{aligned} \tag{B.4}$$

Furthermore, by the choice of α, β and by eq. (B.1) it can be derived that

$$\mathbb{E}[M_\tau] = e^{-\lambda\frac{-a+b}{2}} + e^{\lambda\frac{-a+b}{2}} = 2\cosh\left(\lambda\frac{-a+b}{2}\right). \tag{B.5}$$

Combining eqs. (B.4) and (B.5) gives:

$$\mathbb{E}\big[e^{-\tau\log[\cosh\lambda]}\big] = \frac{\cosh\left(\frac{b-a}{2}\lambda\right)}{\cosh\left(\frac{a+b}{2}\lambda\right)}. \tag{B.6}$$

This is nearly a moment generating function. Such a function is of the following form: $\mathbb{E}[e^{-\theta\tau}]$. Therefore one must substitute $\theta = \log[\cosh\lambda]$. λ can be expressed in terms of θ:

$$\begin{aligned}\theta &= \log(\cosh\lambda) \\ e^\theta &= \cosh\lambda \end{aligned}$$

Using the expression $\cosh^{-1}(x) = \log\left(x + \sqrt{x^2-1}\right)$, one finds:

$$\begin{aligned}\lambda(\theta) &= \log\left(e^\theta + \sqrt{e^{2\theta}-1}\right) \\ &= \log\left(e^\theta\left(1 + \sqrt{1-e^{-2\theta}}\right)\right) \\ &= \theta + \log\left(1 + \sqrt{1-e^{-2\theta}}\right) \end{aligned} \tag{B.7}$$

for $\theta \geq 0$.

Therefore, for the special case of the random walk with $p = q = 0.5$, the moment generating function of the stopping time τ is

$$\mathbb{E}[e^{-\theta\tau}] = \frac{\cosh\left(\frac{b-a}{2}\lambda(\theta)\right)}{\cosh\left(\frac{a+b}{2}\lambda(\theta)\right)} \tag{B.8}$$

with $\lambda(\theta)$ given by eq. (B.7).

C The moment generating function for a Brownian motion with drift

In section 6.3 the moment generating function of the stopping time of a Brownian motion with drift is derived using martingales. In this appendix it is proved that the used martingale is indeed a martingale and the full derivation of the moment generating function is given.

C.1 Proof that M_t is a martingale

Recall the definitions of the random processes M_t and X_t:

$$M_t = \exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t - \mu X_t \right\} \sinh(\lambda X_t - \alpha) \tag{C.1}$$

$$X_t = W_t + \mu t \tag{C.2}$$

with W_t a standard Brownian motion as defined in definition 6.7. First M_t can be rewritten in terms of W_t:

$$M_t = \exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t - \mu W_t - \mu^2 t \right\} \sinh(\lambda W_t + \lambda\mu t - \alpha)$$

Then using the definition of the hyperbolic sine and several calculus rules this result can be rewritten:

$$\begin{aligned}
&= \frac{1}{2} \exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t - \mu W_t - \mu^2 t \right\} \left(e^{\lambda W_t + \lambda\mu t - \alpha} - e^{-\lambda W_t - \lambda\mu t + \alpha} \right) \\
&= \frac{1}{2} \exp\left\{ -\frac{1}{2}(\lambda^2 - 2\lambda\mu + \mu^2)t + (\lambda - \mu)W_t - \alpha \right\} \\
&\quad - \frac{1}{2} \exp\left\{ -\frac{1}{2}(\lambda^2 + 2\lambda\mu + \mu^2)t - (\lambda + \mu)W_t + \alpha \right\} \\
&= \frac{1}{2} \exp\left\{ -\frac{1}{2}(\lambda - \mu)^2 t + (\lambda - \mu)W_t - \alpha \right\} \\
&\quad - \frac{1}{2} \exp\left\{ -\frac{1}{2}(\lambda + \mu)^2 t - (\lambda + \mu)W_t + \alpha \right\}.
\end{aligned} \tag{C.3}$$

Now one can observe that $M_t = \frac{1}{2}N_{+,t} - \frac{1}{2}N_{-,t}$ with $N_{\pm,t}$ defined as

$$N_{\pm,t} = \exp\left\{ -\frac{1}{2}(\lambda \mp \mu)^2 t \pm (\lambda \mp \mu)W_t \mp \alpha \right\}. \tag{C.4}$$

Both M_t and $N_{\pm,t}$ are a function of the random variable W_t. Therefore both random variables are $\mathcal{F}_t$-adapted with $\mathcal{F}_t = \sigma\{W_s | 0 \leq s < t\}$. Then it follows from proposition 6.5 that M_n is a martingale if $N_{\pm,t}$ is a martingale. Therefore it must be proved that the three defining properties of a martingale (definition 6.2) hold for $N_{\pm,t}$.

First we have seen that $N_{\pm,t}$ is $\mathcal{F}_n$-adapted. Therefore the first property holds. Secondly, since W_t is bounded by $-a$ and b, and $t \geq 0$, one can see that $\mathbb{E}[N_{\pm,t}] < \infty$, assuming the constants α, λ, μ are finite.

Finally it must be shown that $\mathbb{E}[N_{\pm,t}|\mathcal{F}_t] = N_{\pm,s}$ for $0 \leq s < t$. It can be seen that

$$\mathbb{E}[N_{\pm,t}|\mathcal{F}_t] = \mathbb{E}\left[\exp\left\{-\frac{1}{2}(\lambda \mp \mu)^2 t \pm (\lambda \mp \mu)W_t \mp \alpha\right\}\middle|\mathcal{F}_t\right].$$

By adding two cancelling terms with s this can rewritten as

$$\begin{aligned}= \mathbb{E}\Bigg[&\exp\left\{-\frac{1}{2}(\lambda \mp \mu)^2(t-s) \pm (\lambda \mp \mu)(W_t - W_s)\right\} \\ &\cdot \exp\left\{-\frac{1}{2}(\lambda \mp \mu)^2 s \pm (\lambda \mp \mu)W_s \mp \alpha\right\}\middle|\mathcal{F}_t\Bigg].\end{aligned}$$

Notice that s and t are deterministic and $W_s \in \mathcal{F}_t$. Therefore several terms can be taken out of the expectation operator:

$$\begin{aligned}= e^{-\frac{1}{2}(\lambda\mp\mu)^2(t-s)} \cdot \mathbb{E}&\left[\exp\left\{\pm(\lambda \mp \mu)(W_t - W_s)\right\}\middle|\mathcal{F}_t\right] \\ &\cdot \exp\left\{-\frac{1}{2}(\lambda \mp \mu)^2 s \pm (\lambda \mp \mu)W_s \mp \alpha\right\}\end{aligned} \tag{C.5}$$

By the third property of definition 6.7, one can find that $(W_t - W_s)$ is normally distributed with mean 0 and variance $(t-s)$. Furthermore it is known that for a random variable X which is normally distributed with mean μ_0 and variance σ^2, the following holds by its moment generating function: [4]

$$\mathbb{E}[e^{tX}] = e^{\mu_0 t} e^{\frac{1}{2}\sigma^2 t^2}$$

Using this property it can be found that

$$\mathbb{E}\left[\exp\left\{\pm(\lambda \mp \mu)(W_t - W_s)\right\}\middle|\mathcal{F}_t\right] = e^{\frac{1}{2}(t-s)(\lambda\mp\mu)^2}. \tag{C.6}$$

Combining this result with eq. (C.5), it can be seen that

$$\mathbb{E}[N_{\pm,t}|\mathcal{F}_t] = \exp\left\{-\frac{1}{2}(\lambda \mp \mu)^2 s \pm (\lambda \mp \mu)W_s \mp \alpha\right\} = N_{\pm,s}. \tag{C.7}$$

Having proved the third defining property of a martingale for $N_{\pm,t}$, it can be said that $N_{\pm,t}$ is a martingale. Therefore, M_t is a martingale as well.

C.2 Derivation of the moment generating function

In this appendix the moment generating function of the stopping time of a Brownian motion with drift is derived. Recall the results from section 6.3. The used martingale is

$$M_t = \exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t - \mu X_t \right\} \sinh(\lambda X_t - \alpha) \tag{C.8}$$

and in eq. (6.16) it was found that $\mathbb{E}[M_0] = \sinh(-\alpha)$. The equation $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$ needs to be solved to find the moment generating function.

First $\mathbb{E}[M_\tau]$ is written out. At time τ the walker is, by the definition of τ, either at position $-a$ or b. Therefore $X_\tau \in \{-a, b\}$. It can be seen that

$$\begin{aligned}\mathbb{E}[M_\tau] = \mathbb{E}\left[\exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t + \mu a \right\} \sinh(-\lambda a - \alpha) I(X_\tau = -a)\right] \\ + \mathbb{E}\left[\exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t - \mu b \right\} \sinh(\lambda b - \alpha) I(X_\tau = b)\right].\end{aligned} \tag{C.9}$$

The parameter $\alpha \in \mathbb{R}$ can be chosen freely. It is chosen such that

$$e^{\mu a} \sinh(-\lambda a - \alpha) = e^{-\mu b} \sinh(\lambda b - \alpha). \tag{C.10}$$

Using the definition of the hyperbolic sine and several calculus rules, eq. (C.10) can be rewritten into the following form:

$$e^{2\alpha} = \frac{e^{(\lambda-\mu)b} - e^{-(\lambda-\mu)a}}{e^{-(\lambda+\mu)b} - e^{(\lambda+\mu)a}}. \tag{C.11}$$

Furthermore since $X_\tau \in \{-a, b\}$, it follows that $I(X_\tau = -a) + I(X_\tau = b) = 1$. Combining this with eq. (C.10), eq. (C.9) can be reduced to the following form:

$$\mathbb{E}[M_\tau] = \mathbb{E}\left[\exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t + \mu a \right\} \sinh(-\lambda a - \alpha)\right]$$

Then according to the results from eq. (6.17) one can see that

$$\begin{aligned}&\mathbb{E}\left[\exp\left\{ -\frac{1}{2}(\lambda^2 - \mu^2)t + \mu a \right\} \sinh(-\lambda a - \alpha)\right] = \sinh(-\alpha) \\ \Rightarrow\ &\mathbb{E}\left[e^{-\frac{1}{2}(\lambda^2-\mu^2)t}\right] = \frac{e^{-\mu a}\sinh(-\alpha)}{\sinh(-\lambda a - \alpha)} = \frac{e^{-\mu a}e^{-\alpha}(1 - e^{2\alpha})}{e^{-\alpha}(e^{-\lambda a} - e^{\lambda a}e^{2\alpha})}\end{aligned} \tag{C.12}$$

Equation (C.12) can be rewritten into a neater result. First the expression for α given in eq. (C.11) is substituted into the equation:

$$\mathbb{E}\left[e^{-\frac{1}{2}(\lambda^2-\mu^2)t}\right] = \frac{e^{-\mu a}(1 - \frac{e^{(\lambda-\mu)b} - e^{-(\lambda-\mu)a}}{e^{-(\lambda+\mu)b} - e^{(\lambda+\mu)a}})}{e^{-\lambda a} - e^{\lambda a}\frac{e^{(\lambda-\mu)b} - e^{-(\lambda-\mu)a}}{e^{-(\lambda+\mu)b} - e^{(\lambda+\mu)a}}}$$

Both the numerator and denominator contain a subfraction with equal denominators. By multiplying both the numerator and denominator of the main fraction by this denominator, the expression simplifies to

$$\mathbb{E}\left[e^{-\frac{1}{2}(\lambda^2-\mu^2)t}\right] = \frac{e^{-\mu a}(e^{-(\lambda+\mu)b} - e^{(\lambda+\mu)a} - e^{(\lambda-\mu)b} + e^{-(\lambda-\mu)a})}{e^{-\lambda a}(e^{-(\lambda+\mu)b} - e^{(\lambda+\mu)a}) - e^{\lambda a}(e^{(\lambda-\mu)b} - e^{-(\lambda-\mu)a})}$$

Working out the brackets reduces the expression to the following:

$$= \frac{-2\sinh(\lambda a) - 2e^{-\mu a-\mu b}\sinh(b\lambda)}{e^{-\mu b}(-e^{\lambda(a+b)} + e^{-\lambda(a+b)})}$$

Multiplying both the numerator and denominator of the fraction by $-e^{\mu b}$ results in a neat expression:

$$= \frac{e^{\mu b}\sinh(\lambda a) + e^{-\mu a}\sinh(\lambda b)}{\sinh(\lambda(a+b))} \tag{C.13}$$

The moment generating function is of the form $\mathbb{E}[e^{-x\tau}]$, therefore the following substitution should be done:

$$x = \frac{1}{2}(\lambda^2 - \mu^2) \Rightarrow \lambda(x) = \pm\sqrt{2x+\mu^2}. \tag{C.14}$$

Note that, since the hyperbolic sine is an odd function, the sign of $\lambda(x)$ can be chosen freely. This can be seen by the following:

$$\begin{aligned}\frac{e^{\mu b}\sinh(-\lambda a) + e^{-\mu a}\sinh(-\lambda b)}{\sinh(-\lambda(a+b))} &= \frac{-e^{\mu b}\sinh(\lambda a) - e^{-\mu a}\sinh(\lambda b)}{\sinh(-\lambda(a+b))} \\ &= \frac{e^{\mu b}\sinh(\lambda a) + e^{-\mu a}\sinh(\lambda b)}{\sinh(\lambda(a+b))}.\end{aligned}$$

Therefore choose $\lambda(x) = \sqrt{2x+\mu^2}$.

Then the moment generating function for the stopping time of a Brownian motion with drift is

$$\mathbb{E}[e^{-x\tau}] = \frac{e^{\mu b}\sinh(a\lambda(x)) + e^{-\mu a}\sinh(b\lambda(x))}{\sinh((a+b)\lambda(x))} \tag{C.15}$$

with $\lambda(x) = \sqrt{2x+\mu^2}$, restricted by $x > 0$ as derived in section 6.2.