



Technische Universität München
Physik-Department, T30g

Facilitated diffusion models for gene regulation in living cells

Maximilian Bauer

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. Friedrich C. Simmel

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Ralf Metzler,
Universität Potsdam
2. Univ.-Prof. Dr. Martin Zacharias

3. Gutachter:

Univ.-Prof. Dr. Udo Seifert,
Universität Stuttgart
(nur schriftliche Beurteilung)

Die Dissertation wurde am 07.08.2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 02.12.2014 angenommen.

Summary

To survive in an ever-changing environment any living organism not only needs to know how to synthesise proteins, but it also needs to be able to judge under which circumstances they should be produced. A single molecule present in any cell, the DNA, contains the blueprints for proteins, yet it also has sites to which other molecules can bind in order to enhance or prevent the production of these proteins. These helper molecules are specialised proteins, called transcription factors. For the survival of a cell it is important that their association reactions with the functional sites on DNA proceed quickly. In the prosaic view of a theoretical physicist this reaction can be simply considered to be a search process, but in fact, this topic is a fascinating example of interdisciplinary research, where biology meets physics and where both fields benefit from findings of the other.

The first chapter of this work presents a historical introduction into the topic, highlighting the central role of the double-helical DNA. It is described how genes are expressed in order to build proteins, and how this expression is regulated. Here the emphasis lies on bacterial cells, since they constitute simpler systems than plants or animals and are often better characterised quantitatively.

Chapter two reviews how theoretical models describe the association reaction of a transcription factor with its target sequence on DNA. Specifically, the so-called facilitated diffusion model whose name appears in the title of this thesis is introduced. Established in the 1970s it explains the experimentally measured high association rates as resulting from a beneficial combination of search phases in the bulk solution and along the DNA molecule.

The last decades saw an enormous progress in experimental techniques. Therefore, the third chapter presents a generalisation of the classical facilitated diffusion model to the current state of scientific knowledge. A general problem in the field is to reconcile the fast motion of proteins along the DNA molecule with their ability to bind tightly to the target site. Therefore we combine a common assumption that the searching protein is present in two conformations with the full classical search model.

While this model successfully describes the situation in *in vitro* experiments, the model introduced in chapter four deals with the core issue of this thesis and directly depicts the search process in a living bacterial cell. Also based on the general concept of the facilitated diffusion model, this semi-analytical approach importantly relies on a coarse-grained description of the bacterial genome.

In the final chapter the real nucleotide sequence of an *E. coli* strain is used to paint a more detailed microscopic picture of the search process. A continuous transition between a model in which the particle switches blindly between its two conformational states and a model in which this interconversion is strongly coupled to the underlying nucleotide sequence is studied. Besides, the presence of other non-specifically bound proteins is explicitly taken into account. Finally, we consider that some proteins are able to bind to two operators simultaneously and loop out the intervening DNA which adds a new layer of complexity to this search problem. Hopefully the models presented in this thesis are steps towards the ultimate goal of a comprehensive understanding of the regulation of prokaryotic gene expression.

List of publications

While working on this PhD project the following papers were published/submitted:

1. Maximilian Bauer and Ralf Metzler, Generalized facilitated diffusion model for DNA-binding proteins with search and recognition states, *Biophysical Journal* **102**, 2321 (2012).
2. Maximilian Bauer and Ralf Metzler, In Vivo Facilitated Diffusion Model, *PLOS ONE* **8**, e53956 (2013).
3. Maximilian Bauer, Aljaž Godec, and Ralf Metzler, Diffusion of finite-size particles in channels with random walls, *Physical Chemistry Chemical Physics* **16**, 6118 (2014).
4. Aljaž Godec, Maximilian Bauer, and Ralf Metzler, Collective dynamics effect transient subdiffusion of inert tracers in gel networks, preprint: arXiv:1403.3910 (submitted).
5. Maximilian Bauer, Emil S. Rasmussen, Michael A. Lomholt, and Ralf Metzler, TF searching for a target in a real sequence, (in preparation).

Contents

1	Introduction	1
1.1	Historical notes: from Plato to Watson and Crick	1
1.1.1	Mendel's experiments	2
1.1.2	Twentieth century: Mendel reloaded	2
1.1.3	Hereditary information: in proteins or in DNA?	4
1.2	The central dogma of molecular biology	6
1.2.1	The structure of DNA	6
1.2.2	RNA and transcription	7
1.2.3	Proteins and translation	8
1.2.4	Scheme of the central dogma	9
1.3	Gene regulation in prokaryotes	11
1.3.1	<i>E. coli</i> and its metabolism	11
1.3.2	Lac operon and its control	12
1.4	There is more than just O_1	14
1.4.1	The symmetric operator O_{sym}	15
1.4.2	The auxiliary operators	15
1.4.3	Structure of the lac repressor	16
1.4.4	Looping	17
1.4.5	Connection to information theory	18
1.5	Sequence specificity and non-specific binding	18
1.5.1	Thermodynamic models	20
2	Biological search processes	23
2.1	The diffusion limit	23
2.1.1	First passage time formalism	24
2.1.2	Experimental results by Riggs et al.	26
2.1.3	Extension of Smoluchowski's formula	26
2.2	Early studies	27
2.2.1	Reduction of dimensionality	27
2.2.2	The contributions of Peter H. Richter	28
2.2.3	The contributions of Otto G. Berg	30
2.3	The Berg-von Hippel or facilitated diffusion model	33
2.3.1	How is sliding made possible biologically?	35
2.4	Modern studies	35
2.4.1	Experimental studies	36
2.4.2	Theoretical studies	37
2.4.3	Computational models	39
2.4.4	Two-state models and the speed-stability paradox	40

2.5	General features of the model and criticism	42
2.5.1	General features	42
2.5.2	Modern criticism	43
3	Generalised facilitated diffusion model	45
3.1	The model	45
3.1.1	3D kernel functions for the straight DNA conformation	49
3.1.2	Solution of the differential equations	51
3.2	Results of GFDM for straight DNA	52
3.2.1	Result obtained with the reference parameter values	52
3.2.2	Dependence on the switching rates k_{rs} and k_{sr}	56
3.2.3	Dependence on the diffusion coefficients	58
3.2.4	Dependence on the reaction volume	60
3.3	Results for coiled DNA	61
3.4	Relation to previously published models	63
3.5	Summary and outlook	67
4	In vivo facilitated diffusion model	69
4.1	Organisation of bacterial DNA	70
4.1.1	Model genome	71
4.2	General search model	72
4.2.1	Details of the search model	72
4.3	Microscopic model	74
4.3.1	Derivation of the target detection probability, p_t	74
4.3.2	Derivation of the non-specific association probability, p_r	77
4.4	Derivation of the mean target search time	80
4.5	Results of the IVFDM	81
4.5.1	Reference set of parameters	81
4.5.2	Bound fraction of time	83
4.5.3	Mean search time	84
4.5.4	Searching at near optimal conditions	85
4.5.5	Influence of different parameter sets	86
4.5.6	Acceleration due to local searches	87
4.6	Summary and outlook	89
5	Target search in a real sequence	91
5.1	Search in the target region	92
5.1.1	Score matrix	93
5.1.2	Relation between scores and energies	94
5.1.3	Reference set of parameters	95
5.1.4	Theoretical model	95
5.2	Results for the search in the target region	99
5.2.1	Probability to detect the target	99
5.2.2	Impedance matching	101
5.2.3	Conditional target detection time	102
5.2.4	Probability of first detecting $O1$	103

5.3	Full search model	104
5.4	Results for the full search model	106
5.4.1	Dependence on α	106
5.4.2	Looping effects	108
5.4.3	Blocker conformation effects	109
5.5	Summary	110
6	Discussion and outlook	111

1 Introduction

1.1 Historical notes: from Plato to Watson and Crick

More than two thousand years ago Plato wrote his famous work “state” which is mostly concerned with the design of an ideal state. However, it also contains the following thoughts on heredity [1]:

ἄτε οὖν συγγενεῖς ὄντες πάντες τὸ μὲν πολὺ ὁμοίους ἂν ὑμῖν αὐτοῖς γεννῶτε, ἔστι δ' ὅτε ἐκ χρυσοῦ γεννηθεῖη ἂν ἀργυροῦν καὶ ἐξ ἀργύρου χρυσοῦν ἔκγονον καὶ τᾶλλα πάντα οὕτως ἐξ ἀλλήλων.

Following the translation to German by the Danish botanist Wilhelm Johannsen, Plato states [2]: *As you are all related to each other, you will mostly have descendants which are similar to you; sometimes, however, a silver one can derive from a gold one and vice versa, and similarly in all others.*

Obviously, he could not know that in our times the term “gene” is familiar to almost everyone. But in fact it is derived from the ancient Greek word, which is framed in blue in the above quotation and which can be transliterated as *gennote*¹. Literally it is translated as *generating*². But this excerpt is also interesting because of its notions on which traits are inherited from parents and also on what in modern times might be called mutations. Here the precious metals “silver” and “gold” are metaphors for noble character traits in humans. However, as Johannsen stated, the philosophy of ancient Greece was much more evolved than the actual scientific knowledge at their time. In particular, the notion of genes did not exist. Thus, Plato did not distinguish between the “nature”, i.e. inner traits or in modern terms the “genotype” and external stimuli which change the outer appearance, the “phenotype” [2].

Nowadays, due to widespread use in crime thrillers or in forensic science most people have a notion on what DNA is. Besides, the word gene can often be found in newspapers, be it in the context of genetically modified food and tests for hereditary diseases or even on sports pages when it is discussed whether or not a team possesses a “winner gene”. These terms appear natural to our contemporary ears, but it is amazing to recapitulate how few was known scientifically about this topic 150 years ago, when the friar Gregor

¹Interestingly, it was Johannsen who coined the term “gene” [3]. One may speculate if it was this very excerpt which motivated this choice.

²Actually, the word within the red frame can be transliterated as *allelon* and is translated as *apart*. It is the root for the word “allele” which is an important technical term in genetics as well

Johann Mendel conducted his “experiments on plant hybridization” whose results were published in 1866 [4].

1.1.1 Mendel’s experiments

Obviously, just like Plato Mendel could not know what a gene is. But the hereditary units which he called *factors*, are in fact genes. In commenting on the importance of his findings, we follow the description of Ilona Miko [5]. Mendel’s choice to breed *pisum sativum*, colloquially known as pea plants, was clever because they can be both self-fertilized and cross-fertilized [5]. Only this versatility enabled him to reach his conclusions. Another important point in his studies was that he focused on seven traits of the pea plants, which each could attain only two “values”. For example, he studied the pod shape, which could be either constricted or inflated [5]. This binary form enabled an analytical or even mathematical description of the results.

His main interpretation was that *factors* which are responsible for the occurrence of visible traits, are inherited from both parents. Factors can be present in different variations and therefore it is possible that an organisms receives different sets of factors from its parents. Importantly, he introduced the concept that these alternative variations of a factor can be dominant or recessive. Therefore he studied what the progeny looks like when their parents share all traits but one [5]. The result is usually that concerning the differing trait the offspring will not look like a blend of their parents but according to the dominant trait.

However, in the second half of the nineteenth century no one could expect that nowadays many people consider him the forefather of genetics. While Mendel’s observations were truly ahead of his time, what kind of substance hosts these factors remained unknown. Accordingly, his work only started being appreciated in the twentieth century, when further advances in experimental techniques were made. Thus, it is no surprise that no one noticed the relation between Mendel’s results and the ones of his contemporary, the Swiss biologist Johannes Friedrich Miescher. He studied leukocytes in the pus³ of bandages which he obtained from a surgical clinic in Tübingen [3]. While doing this, he isolated the substance “nuclein”—which later was identified as DNA, i.e. the carrier of Mendelian factors—for the first time in 1869 and published these findings in 1871 [3].

Less than hundred years later, in 1953, James D. Watson and Francis H. C. Crick published an article entitled “Molecular Structure of Nucleic Acids - A Structure for Deoxyribose Nucleic Acid” [6]. In this article they suggested that DNA has a double-helical structure. We will now review which scientific findings happened between these two events, where we follow the description presented by Ralf Dahm [3].

1.1.2 Twentieth century: Mendel reloaded

Nearly three decades had to pass after Mendel’s findings until around 1900 several scientists, namely Carl Correns, Hugo de Vries and Erich von Tschermak rediscovered them [3]. Already in 1902, the American physician Walter S. Sutton wrote the following

³The yellow or white substance found for example in an abscess.

sentence in an article entitled “On the morphology of the chromosome group in *Brachystola magna*⁴” [7]:

“I may finally call attention to the probability that the association of paternal and maternal chromosomes in pairs and their subsequent separation during the reducing division as indicated above may constitute the physical basis of the Mendelian law of heredity.”

Thus, Sutton and likewise his German contemporary, Theodor Boveri established a connection between Mendel’s theory and research on chromosomes [3]. These chromosomes had been found by the German biologist Walther Flemming several years after Miescher had observed “nuclein”. Flemming’s experiments had yielded that cell nuclei contain fibrous networks and he was even able to describe their motion during cell division [3].

In the following years, it was common knowledge among scientists that the Mendelian factors which were by now called genes are to be found on chromosomes and that the main constituents of these are proteins and DNA [8]. But among these two, proteins were thought to be the more suitable choice for storing genetic information, since their chemical and physical structure is far more complex than the one of DNA. However, this assumption was proven wrong in a series of experiments as we will recapitulate now.

In the 1920s Frederick Griffith worked with two strains of *streptococcus pneumoniae*, a bacterium which—as its name already implies—causes pneumonia [9]. The r strain (where r represents rough) was found to be less pathogenic than the s strain (where s stands for smooth), which is covered with a protective capsule of polysaccharides [8]. If bacteria of the dangerous s strain were killed by heating them up, their injection into mice did not affect their health. If, however, these dead bacteria were injected alongside living bacteria of the less dangerous r strain, the mice died. This surprising finding could only be rationalised if one assumes that a “transforming principle” is at work. This means that the r strain obtained the ability, i.e. was transformed, to form the protective capsule from some component of the s strain which was not destroyed when the s strain bacterium was killed by the heat [8]. But it was not clear how exactly to interpret these observations and the most important question was: what exactly is the “transforming principle”?

It took another two seminal experiments to answer this. The first one was performed by Oswald T. Avery, Colin MacLeod and Maclyn McCarty in the middle of the 1940s. It was a modification and extension of Griffith’s earlier experiment, this time, however, performed in cultures [8, 10]. They used an exclusion principle, in which the original set of s strains which had previously been heated was separated in three parts. The first one was subsequently treated with DNase, the second one with protease and the third one with RNase. These substances with the suffix “-ase” are known to destroy the corresponding substances. Then the reaction products were again mixed with r strain cells. From the observation that the DNase treated s cells did not transform the r strain, while the other two substances still had this ability, they deduced that DNA is essential for the transformation.

In his influential book “What is life” the Austrian physicist Erwin Schrödinger also reasoned about the role of chromosomes as an “hereditary code-script” [11]. Importantly, he pointed out that chromosomes are more than just a code-script [11]:

⁴*Brachystola magna* refers to a type of grasshoppers.

“The chromosome structures are at the same time instrumental in bringing about the development they foreshadow. They are law-code and executive power - or, to use another simile, they are architect’s plan and builder’s craft - in one.”

Besides, by invoking an analogy to the Morse code, where a small set of letters is sufficient to write complex texts, he advocated the idea that molecules with a rather simple structure can be responsible for the formation of more complex molecules [11].

1.1.3 Hereditary information: in proteins or in DNA?

Still there were scientists who favoured the protein to be the carrier of hereditary information. The straw that broke the camel’s back was the second seminal experiment mentioned above which was performed by Alfred Hershey and Martha Chase in 1952 [12]. They studied the phage⁵ T2 which infects the bacterium *Escherichia coli* (*E. coli*), named after the paediatrician Theodor Escherich, and exploited the fact that proteins and DNA differ chemically: namely, unlike proteins DNA contains phosphate and conversely proteins have a sulphurous content, but not DNA [8]. Using radioactive isotopes of these two elements, Hershey and Chase were able to show that the protein simply forms a coat around the phage while its DNA is injected into the bacterium. Knowing in retrospect how important their findings were, it is intriguing to see how cautiously they summarised their findings, stating simply that “the DNA has some function” [8, 12].

One year before Oswald T. Avery and co-workers published their work, Max Delbrück and Salvador E. Luria who later became the doctoral advisor of James D. Watson answered a question which bothered scientists since the 1920s. Namely whether the immunity of some *E. coli* cells to a bacteriophage results from random mutations or are a direct consequence of the interaction with the virus [14]. Backed up by a mathematical theory they could show that the distribution of survivors they found experimentally was not Poissonian and could only be rationalised by assuming that the mutations happened randomly and thus independently of the presence of the phage.

In the time between the experiments of Avery and of Hershey, the Austrian biochemist Erwin Chargaff studied the base composition of DNA in more detail. While it was known that DNA contains the two purines adenine (A) and guanine (G) and the two pyrimidines cytosine (C) and thymine (T), for a long time wrong conclusions were drawn. In particular, the biochemist Phoebus Levene had formulated the “tetranucleotide theory”: within this theory, DNA which was called yeast nucleic acid by Levene, was supposed to be build up of repeating units of these four bases. This implies that in any species all four bases should appear equally often. Chargaff, however, was able to show two points: that while every species may have a characteristic frequency with which the four bases occur, the frequency of As is equal to the one of Ts. An analogous rule applies for Cs and Gs [3].

Similarly important was the finding of George Beadle and Edward Tatum published in an article in 1941, where they reasoned that genes “control or regulate specific reactions

⁵Nearly as important for molecular biology as bacteria themselves are the viruses that infect them, the (bacterio-)phages. Literally, this means *bacteria eaters*. A whole group of biologically interested physicists, most notably Max Delbrück, was named phage group after them [13]. We will encounter Delbrück’s most important contribution to the topic of this thesis in section 2.2.

in the system either by acting directly as enzymes or by determining the specificities of enzymes” [3, 15]. A statement which later was shortened to the catchy slogan: “one gene-one enzyme hypothesis”. Furthermore, in 1949 Colette and Roger Vendrely and André Boivin proved that somatic cells contain twice times the amount of the one found in germ cells [3].

The importance of the scientific findings obtained in the first half of the 1950s cannot be overstated: only one year after the experiment of Hershey and Chase, the renowned journal *Nature* featured two articles on the structure of DNA in a single issue. The first one written by Rosalind Franklin and Maurice Wilkins described X-ray studies of DNA and contained the by-now legendary “photo 51” [3]. The second one by Crick and Watson was the one already mentioned which is arguably one of the most important scientific publications of the last century [6]. It set the base of what is known about the structure of DNA and thus enabled scientists to decipher the genetic code and to understand how gene expression works. This will be discussed in the following section, but first we make a few general remarks.

Living organisms are usually classified as belonging to one of the following three classes or *domains*: (eu)bacteria, archae(bacteri)a and eukaryotes. The first two domains together are called prokaryotes and unlike eukaryotes they do not contain a nucleus⁶, i.e. a compartment of the cell which is surrounded by membranes and contains the DNA [13]. The cell nucleus was discovered by Robert Brown. This work focuses on gene regulation in prokaryotes. The reason for that being that they constitute simpler systems compared to higher organisms. From a technical point of view it plays a role that they usually have short doubling times, easing the experimental analysis [16].

Out of all bacteria, the most prominent example is *E. coli*, which was already mentioned above. Mostly due to the work of Jacques Monod and co-workers its metabolism came to the centre of attention. In 1957 Aaron Novick and Milton Weiner published their seminal work showing that the induction of the enzyme β -galactosidase is an “all-or-none” phenomenon [17]. This means that a colony of bacteria is very heterogeneous concerning the rate at which individuals produce this enzyme. Some do this at full throttle, while others nearly not at all. Such a behaviour is also called bistable, where the prefix “bi-” indicates that there are two stable states. They further hypothesised that the critical step for induction is the formation of a single specific enzyme. However, this hypothesis has recently been tested and disproved using modern single molecule techniques [18].

This closes our historical introduction which started in ancient Greece and in which we encountered diverse living organisms ranging from pea plants, over grasshoppers to the bacterium *E. coli*. For roughly six decades we know that DNA plays *the* central role in cellular biology, but in the following we include two other types of molecules into our considerations. Thus, we focus our attention on the three most important classes of biopolymers in a cell: DNA, ribonucleic acid (RNA) and proteins.

The following sections are ordered rather topically than historically. In order to understand the importance of DNA, RNA and proteins, we study their structure, starting with DNA and subsequently we consider how the expression of genes is regulated. This will enable us to understand the so-called central dogma of molecular biology.

⁶They do, however, possess a similar object called *nucleoid*, which will be considered in chapter 4.

1.2 The central dogma of molecular biology

In describing the structure and function of these essential biopolymers we follow the book “Molecular biology of the cell” and Robijn F. Bruinsma’s review article “Physics of protein-DNA interaction” [13, 19]. The central dogma was introduced by Francis H. C. Crick who discussed how sequential information is transferred [20].

1.2.1 The structure of DNA

One might naively ask why the structure of a particular biopolymer is of such relevance, especially given that DNA in itself is not even very reactive [16]. The answer to this is that this special structure enables the storage of information. All living organisms, be it bacteria, plants or animals, store not less than their hereditary information in DNA [13].

In most general terms, one can say that DNA consists of two strands which are linear polymer chains. These are composed of monomers, the nucleotides. In turn, a nucleotide is composed of a backbone made up of the pentose sugar deoxyribose with an tetrahedral phosphate group, PO_4^{3-} , attached to it which connects neighbouring sugars [13, 19]. Roughly, the backbones can be considered to be equivalent to the stringers of a ladder.

Most importantly the nucleotide contains one of the following four bases: the two purines adenine (A) and guanine (G) or the two pyrimidines cytosine (C) and thymine (T). These bases build the rungs of the ladder. Since the backbone is the same for all monomers, the content of a single strand is completely determined by the sequence of A, C, G and T. However, this also fixes the composition of the second strand via the complementary rule that an A on one strand builds a base pair⁷ with a T on the other strand (and vice versa). The same applies for the two bases C and G. Thus, a more spacious purine with two rings always pairs with a smaller pyrimidine that has just one ring. The complementary bases are connected via hydrogen bonds, two in the case of A and T and three between C and G [19]. The complementarity is often compared to a lock-and-key mechanism and is important for polymerase chain reactions (PCR), for example when crime scenes are investigated.

The second force stabilizing the DNA is the stacking interaction, a hydrophobic attraction between bases [19]. The picture of a ladder-like DNA introduced above is however, too simple a picture, since the two strands twist around one another to form the double helix that was predicted by Watson and Crick based on Franklin and Wilkins’ observations [6] (compare Fig. 1.1).

Due to the way the two helices twist around each other two differently sized grooves emerge, which are called *major* and *minor groove*. Finally, it is noteworthy that the backbone has a directionality: by convention, the carbon atoms in the sugar are numbered. Of particular importance are the third and the fifth carbon atom which roughly denote the orientation of the stringer of the symbolic ladder. Thus, one speaks of the 5’-end where the chain ends with the carbon atom number five of the sugar. And accordingly, the 3’-end consists of the hydroxyl group which is attached to the third carbon atom. This directionality has interesting physical consequences: the differences in driving a single-stranded DNA in either direction through an α -hemolysin pore can be explained

⁷Actually, the term base pair is also used to designate the length that such a unit spans on DNA. In common units it corresponds to 0.34 nm.



Jmol

Figure 1.1: The double-helical structure of DNA (green and blue) with the dimeric lac repressor DNA-binding domain (red) attached to it. The image was created with the software Jmol, based on the structure (with ID 2KEI) deposited in the Protein Data Bank (PDB) by Romanuka et al. [21, 22]. Note by comparing the pairing bases on the blue and on the green strand that a smaller pyrimidine on one strand always pairs with a large purine on the other strand.

by analogy to the way a tree is brought through a door [23]. In both cases, one direction is clearly preferred.

In a simplistic view one can say that the hereditary information in DNA is written in a four-letter alphabet. But what exactly is this information and how is it read out and used? This is where two other types of biopolymers besides DNA come into play, which are important for the survival of a cell: ribonucleic acid (RNA) and proteins.

Most of the information stored in DNA are blueprints of how to build proteins. In fact, the part of the DNA where the blueprint for a specific protein is written down is called a *gene*⁸. A typical bacterial gene has a length of around 1000 bp, while eukaryotic genes can be much longer. Accordingly, the process in which the protein corresponding to a certain gene is produced, is called gene expression. However, proteins are not built directly from DNA. In technical terms, the information has to be *transcribed* and subsequently it has to be *translated*. Both processes can be described as templated polymerisation [13]. What is meant by this will become clear in the following section.

1.2.2 RNA and transcription

The product of transcription, RNA, has a structure which is very similar to the one of a single DNA strand. It is a linear polymer which involves the sugar ribose (to which its

⁸In some case already the RNA constructed from the DNA sequence is the final product. This type of RNA is called non-coding RNA (ncRNA) and the corresponding DNA segments are called (RNA) genes.

full name *ribonucleic acid* is due) instead of deoxyribose. As implied by their denotations the sugars ribose and deoxyribose differ in the presence of an oxygen atom [19]. Besides, instead of thymine (T) RNA features uracil (U), in which a hydrogen atom is replaced by a methyl group. In general, RNA is much less stable than DNA.

The structural similarity between DNA and RNA enables a more or less straightforward transmission of the information contained in the DNA nucleotide sequence to the one in RNA⁹. The transcription of the text written in DNA language to the one in RNA language is performed by an enzyme called RNA polymerase (RNAP). In bacteria, there is only one type of RNAP, while in eukaryotes there are several ones. It has to bind to a promoter which marks a point on DNA from which RNA synthesis is supposed to start [13]. In order to bind to the promoter the RNAP has to be able to recognise the sequence. This general problem of sequence specificity will be dealt with in a part of the following section. Typically, promoter sequences are not symmetric, thus implicitly telling the RNAP which of the two strands is to be read out [13]. Actually, the design of the promoter is one of the earliest stages at which the cell can regulate gene expression (see section 1.3) [13].

After docking to the DNA, by help of other proteins the RNAP opens the double helix and unwinds it in order to lay open the base pairs [13]. Then the templated polymerisation takes place: one nucleotide after the other is assembled into the transcript, using the DNA sequence as a template where the complementarity of nucleotides helps to make this process called *elongation* nearly error-free. This continues at a rate of approximately 50 nucleotides per second until a *terminator* is encountered on the DNA [13]. This confers a stop signal which makes the RNAP release the RNA transcript. In the “normal” case the RNA which was created is called messenger RNA (mRNA), because its task is simply to contain the information which protein has to be built in the subsequent translation process. However, as already mentioned some of the RNAs produced like this are already functional end products.

The part of DNA which is read out to produce mRNA is called *transcription unit* and can contain one or several genes. Therefore depending on whether it codes for a single or more proteins, one speaks of monocistronic or polycistronic mRNAs [13]. Many copies of mRNA can be produced in a row and sometimes work on a new transcript begins before the previous one is actually finished [13].

1.2.3 Proteins and translation

Translation is a more difficult process, again in the form of a templated polymerisation. This time the mRNA is the template and the products are proteins. They actually perform the jobs to do. In bacteria, they are often produced in bursts [24, 25]. From a structural point of view proteins are again linear polymers whose monomers are α -amino acids collected from a larger alphabet. The monomers are also called residues.

The genetic code The relation between the DNA language and the resulting sequence of amino acids in a protein, is called the *genetic code*. Already in 1961, Francis H. C.

⁹We note that the *RNA world hypothesis* states that in early evolutionary times—before DNA assumed the central role—RNA itself was the bearer of genetic information and induced cellular chemical reactions [13].

Crick and co-workers demonstrated that three consecutive nucleotides (a *codon*) in RNA determine which amino acid is to be included in the protein [26]. In the same year, the biochemist Marshall W. Nirenberg cracked the first codon by showing that RNA solely consisting of uracil will be translated to a protein which is built up exclusively of phenylalanine [27]. Only four years later, the whole genetic code was cracked, see e.g. table 4 in [28], an achievement for which the main scientists in the field received the Nobel Prize in Physiology or Medicine another three years later.

In principle, having a codon consisting of three nucleotides which are taken from a pool of four, implies that there is a total number of $4^3 = 64$ possible combinations. However, under normal circumstances there are less than these 64 theoretically possible different amino acids. Only twenty proteinogenic amino acids are encoded, showing that there is a substantial redundancy. Conversely, it is easily understood that codons of length 2 would yield only $4^2 = 16$ possible combinations and could not explain the presence of the 20 natural amino acids.

Usually, translation is started at a codon which consists of the three bases AUG and stops at one of the three stop codons: UAG, UGA or UAA. Very recently it was shown that modifications in the stop codon are encountered more often in the wild than one would naïvely expect [29].

Structure and function of proteins Proteins typically consist of 50 - 2000 amino acids [13]. For example, the lac repressor which will play a central role in this thesis, has a length of 360 amino acids and of approximately 10 nm in real space [30]. Proteins all share the ability to bind to certain other molecules, the ligands, via a reactive portion of their surface which is called binding site [13]. Such a ligand can be DNA, in which case the protein is called DNA-binding protein (DNABP). There are plenty of reasons why a protein should bind to another molecule like DNA. One which is very important for this work is to prevent that *another* molecule binds to the same or a nearby position. How exactly this happens will be detailed in section 1.5.

Summarising the last two sections one can say that if a protein is to be built the information which amino acids need to be produced and in which order is written in the DNA's nucleotide sequence. The actual process of building a protein is then divided into two steps, transcription and translation. The first refers to the construction of a specific type of RNA called messenger RNA (mRNA) from the DNA template whereas translation refers to the process in which the mRNA is used to build proteins.

Now, we have met the three most important classes of biopolymers in a cell and we have seen how in the usual case the information contained in the nucleotide sequence of DNA is transferred to a sequence of amino acids in a protein. One might be tempted to ask if this transfer of information proceeds on a one way street or if proteins can also influence the information content of DNA. This question is answered by the so-called central dogma of molecular biology.

1.2.4 Scheme of the central dogma

Given that DNA, RNA and proteins are all linear polymers whose monomers are taken from a fixed set, the sequence in which the monomers are present can be considered as a text written in one of the three corresponding languages.

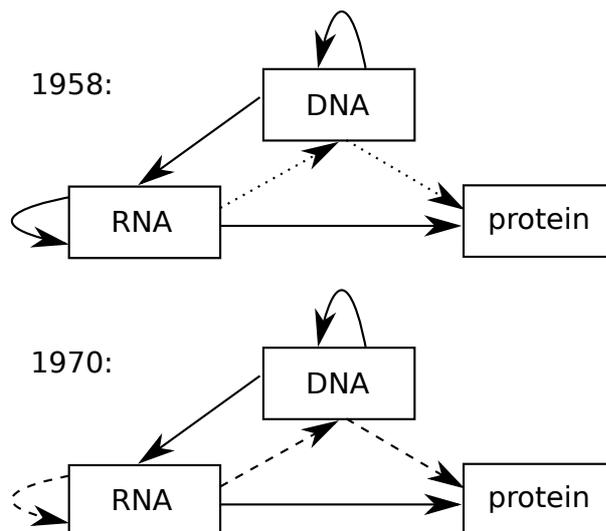


Figure 1.2: Schematic illustration of the central dogma in the original version of 1958 (upper panel) and in its refined version of 1970 (adapted from [20]). In the upper panel full lines correspond to “probable” processes, and dotted lines to “possible” processes. In the lower panel “general” transfers are presented with full lines and “special” transfers are presented with dashed lines [20]. The “impossible” or “unknown” transfers starting from protein are not shown.

Now, in principle there are nine possible ways in which the information conveyed in such a text can be communicated to one of the other two forms of bio-polymers or to another representative of the same class. However, the so-called central dogma of molecular biology states that “once (sequential) information has passed into protein it cannot get out again” [20]. This is schematically illustrated in Fig. 1.2, where both in the original picture of 1958 (upper panel) and in the refined version of 1970 (lower panel) the three processes forbidden by the central dogma are not shown.

The other six processes can be further classified in two groups: in 1958 it was thought that four of them are “probable“ (shown with full lines in the upper panel of Fig. 1.2) and the two others were estimated to be “possible“ (dotted lines). We focus on the refined model of 1970 which is shown in the lower panel of Fig. 1.2 and in which the same six processes are grouped slightly differently: the three which are depicted in dashed lines are only observed under specific conditions, whereas the three processes shown with continuous lines are the ones which were introduced as transcription and translation in the preceding section, plus DNA replication denoted by the arrow starting and ending at DNA.

Given that DNA contains the information about all the proteins that a living organism will produce in its lifetime, it obviously makes sense that it is attempted to keep the information content of DNA from outer influence. Reverse transcription, that is information transfer from RNA to DNA is the usually unwanted exception from the rule, performed by retroviruses. Ironically, many findings on DNA are due to modifying the information content of DNA, for example by irradiating them.

In general, the central dogma should not be taken as dogmatic as it sounds and in the last decades more and more exceptions from the rule were found. However, since we solely study regulation of “normal” transcription, they will not be discussed further here.

1.3 Gene regulation in prokaryotes

In general it is customary to say that a gene is *on*, if it is currently being expressed, and *off* if this is not the case. Since every single cell of a specific organism contains the same DNA, in principle all genes could be *on* at all times. But this is not what happens, rather they are produced only when needed. Such a need can be variable for different cells in an organism, for example a cell in the gut has to behave differently from a cell in a muscle. In fact, some bacterial genes are always produced at a basal rate, which is not the case for eukaryotes. Besides, depending on temperature, the food supply or other external stimuli different genes need to be expressed. This selective use of genes is called regulation of gene expression or shorter gene regulation [16]. It will be considered in this section.

Given that the production of proteins is a process composed of multiple steps, it is clear that gene expression can be switched on and off, or in other words regulated at various stages. The most important class of regulation of gene expression is called *transcriptional control* [13]. In this case already at the stage of transcription it is decided when a gene is expressed and if so at what rate this occurs [13]. This guarantees that no semi-manufactured products are being built that are not needed by the cell, avoiding unnecessary energy costs [13].

Without involving further molecules, as already mentioned the design of a promoter is the first stage where differences in the expression of genes occur. Since not all promoters have the same sequence, it is obvious that the genes whose promoters are “stronger“ are more likely to be bound and expressed by RNAP [13]. Thus, if there are several organisms which differ in the promoter sequence for a gene whose product is often needed, the ones with the stronger promoter have an evolutionary advantage. But before we see in a particular example how transcriptional control works, we introduce in more detail the arguably most important organism for molecular biology.

1.3.1 *E. coli* and its metabolism

The importance of the bacterium *E. coli* was already highlighted in the introduction. In fact, most of our knowledge about the microscopic basics of life stem from studies of this bacterium. Ultimately, the hope is that what is found out to be correct about *E. coli* might enable scientists to draw conclusions for higher organisms, too.

One of the main characteristics of life is that living organisms metabolise. Likewise it is obvious that these organisms are favoured which are able to control their metabolism efficiently. *E. coli* is an organotrophic bacterium, which means that it lives on organic compounds such as lactose and glucose. In order to understand why the preferred choice for *E. coli* is glucose we have to consider their chemical forms.

Lactose, which is colloquially known as milk sugar, is a disaccharide characterised by the formula $C_{12}H_{22}O_{11}$. The prefix “di-“ implies that it is made up of two components

which are in this case the two monosaccharides galactose and glucose, which both have the composition, $C_6H_{12}O_6$. In fact, both sugars are epimers, differing in just one stereogenic centre. To form lactose they are linked via a $\beta - 1,4$ glycosidic bond [31]. Thus, from its chemical form it is obvious that in the presence of glucose it would be wasteful to produce enzymes which cleave lactose since the glucose which results from this can already be used directly.

However, if there is no glucose around three proteins are involved in the metabolism of lactose: β -galactoside permease which is usually bound to the cell membrane, and which imports lactose from the surrounding medium into the cell. Besides, β -galactosidase cleaves the β -1,4 glycosidic bond, thus degrading lactose into its constituents glucose and galactose [31]. The function of the third protein, β -galactoside transacetylase, is less clear [32]. In general, it has a detoxifying function: it acetylates sugars which cannot be metabolised, thus precluding their return into the cell [33, 34]. The genes which encode these proteins are called *lacY*, *lacA* and *lacZ*. Here we follow the convention that genes in bacteria are described by a lower-case italic symbol consisting of three letters, followed by a italic capital [35]. The proteins they code for are written with a capitalised first letter, i.e. LacY, LacA and LacZ.

Since at least two of these three proteins are required to metabolise lactose, it is necessary that they are produced together when needed. This is ensured, since they are put together in an *operon*. What is meant by this will be explained in the following subsection.

1.3.2 Lac operon and its control

Saying that several genes belong to an operon means that they are adjacent to each other and that they share a single promoter. Thus, their corresponding proteins are built in one go. In the paradigmatic case of the lac operon these are the genes coding for the three proteins mentioned in the last subsection. Remembering their tasks it is obvious that such an arrangement makes sense, since there is no point, e.g. to produce an enzyme that imports lactose into the cell if there is no protein to digest it.

However, there is more to an operon than just the promoter and the genes. Even without detailed biological knowledge it is obvious that other molecules can either help or prevent the RNAP from transcribing a gene. An important observation in order to understand this, is that not all of the information on DNA yields blueprints for proteins or functional RNA. In between genes there are stretches of regulatory DNA, i.e. positions where specialised proteins bind to in order to influence the rate at which transcription occurs.

In general there is a plethora of possible ways how to design these regulatory regions on the DNA. We first focus on the case, when a specialised protein, which is called transcription factor (TF) prevents the expression of a gene. These TFs are repressors, since they repress the expression and the regulatory regions which they bind to are called *operators*. They also form a part of the operon and they are typically placed such that if the repressor binds to them, it is not possible for RNAP to bind to the corresponding promoter. This then prevents the initiation of gene expression, making regulation by repressors a type of *negative* regulation. In the case of the lac operon this TF is the lac repressor, LacI and the (main) operator is called *O1*. Its nucleotide sequence is given in

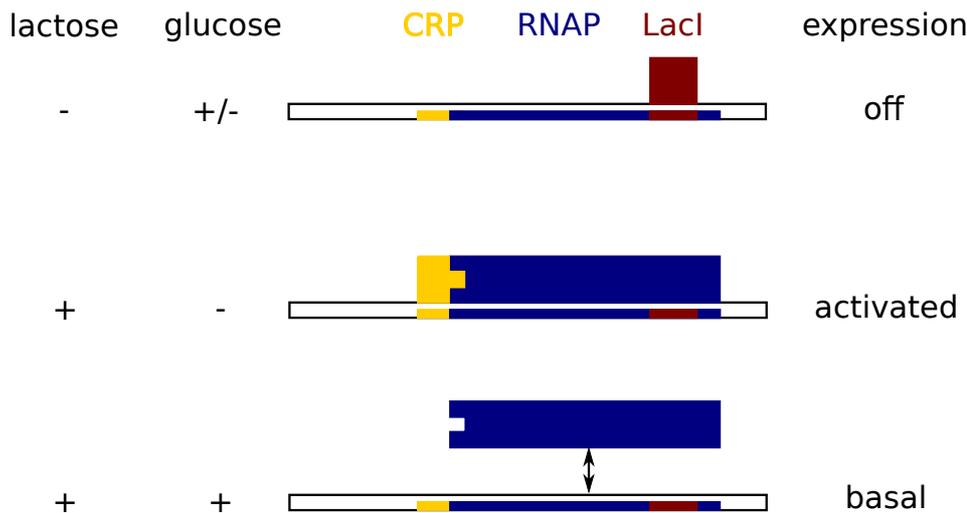


Figure 1.3: Schematic depicting the state of the lac operon depending on the concentration of lactose and glucose (adapted from [38]). CRP (yellow), RNAP (blue) and the lac repressor (dark red) can bind to their respective sites. Specific binding of the repressor shuts off the expression, while RNAP binding starts the expression either in an activated fashion when CRP binds as well or at a basal rate if this is not the case.

the first row of Fig. 1.4 on page 15. The binding of lac repressor to its operator is one example for *specific* binding, since it is an event in which the protein binds to a particular sequence to perform a specific task. This is in contrast with *non-specific* binding which will be detailed below.

It is interesting to note that in the middle of the 1960s it was not even clear what kind of a molecule the operator is. It was only shown in 1967 by Gilbert and Müller-Hill that it is in fact part of the DNA [36]. One year before that the same authors proved several points about the lac repressor: that it is a protein whose gene lies outside the lac operon and that it is present in low copy numbers of approximately 10 per genome [37].

However, the lac operon is also positively regulated by the catabolite activator protein (abbreviated as CAP or CRP). In general this TF helps bacteria to use carbon sources apart from the preferred choice glucose [13]. To fulfil this task, CRP has to bind to cyclic adenosine monophosphate (cAMP). This enables it to bind to the DNA near the promoter, where it acts as an activator [13].

Obviously, it would be a waste of resources if lactose-digesting enzymes were produced in the presence of the favoured glucose. Thus, whenever glucose is present, the concentration of lactose is unimportant and the operon should be shut off. Conversely, if neither glucose nor lactose are present, there is no need to produce the enzymes, too. In other words, the operon should be expressed only when two conditions are met: lactose is present and glucose is absent from the cell [13]. Otherwise it is the task of the repressor to prevent the expression of the lac operon. In other words, interpreting high lactose concentration and low glucose concentration as signals, in terms of logical operations the lac operon plays the role of an AND gate [31]. This is schematically shown in Fig. 1.3.

If lactose is not present in the cell, the lac repressor should bind to its operator to shut off the expression of the lac operon irrespective of whether or not glucose is present (first row in Fig. 1.3). The second row shows the situation when the conditions of the AND gate are met: lactose is present, but not glucose and the lactose-digesting enzymes should be produced. Then, CRP *and* RNAP bind specifically and activated expression of the operon occurs. Finally, when both sugars are present (last row of Fig. 1.3) the lac operon is expressed at some basal rate since even without the help of CRP an RNAP molecule can bind to the promoter to start expression. However, since the concentrations of sugars in the environment usually are not constant, the system must be able to respond to changes.

Response to changes in the environment In order to sense whether the lac operon should be on or off both TFs, the lac repressor and CRP have an activity which depends on the concentration of environmental sugar molecules.

The molecule allolactose is an intermediate metabolite of lactose [39]. Thus, its presence implies that lactose is present in the environment. Allolactose is able to bind to the lac repressor and if so, it reduces the repressor's affinity for the operator [13]. Accordingly, it dissociates and allolactose acts as an inducer of the operon.

The activity of the positive regulator, CRP, is modulated according to the concentration of glucose. This happens indirectly via its activator, cAMP. Usually, cAMP is produced from ATP by adenylate cyclase. However, if glucose is present, this substance is inhibited. Thus, if the glucose concentration increases, the concentration of cAMP in the cell decreases [13]. Then, there are not enough cAMP molecules to bind to CAP. This reduces its affinity for DNA and the positive regulation stops [13]. Conversely, the presence of cAMP conveys the cell that there is a glucose shortage [40].

It is important that there is a positive feedback in this system: when LacY, the permease, is expressed this facilitates the uptake of lactose, which deactivates the repressor, thus further increasing the production of permease [18]. This positive feedback leads to the all-or-none phenomenon that was found by Novick and Weiner [17].

Lactose analogues It was noticed pretty early that not only lactose acts as an inducer for the lac operon. Rather there are lactose analogues which can have experimental advantages. Already in 1957 did Aaron Novick and Milton Weiner notice that methyl- β -D-thiogalactoside (TMG) can be used as an inducer instead of lactose [17]. It is an "gratuitous" inducer, meaning that it is not metabolised by the bacterium. This term was introduced by Jacques Monod [41]. Such a gratuitous inducer greatly facilitates the experimental treatment. Another relevant lactose analogue is isopropyl β -D-1-thiogalactopyranoside (IPTG) [39].

1.4 There is more than just *O*1

Even though, the preceding sections give a correct impression of the main ingredients of the lac system, in reality the situation has even more layers. First of all, the operator *O*1 does not describe the only nucleotide sequence to which the lac repressor can bind.

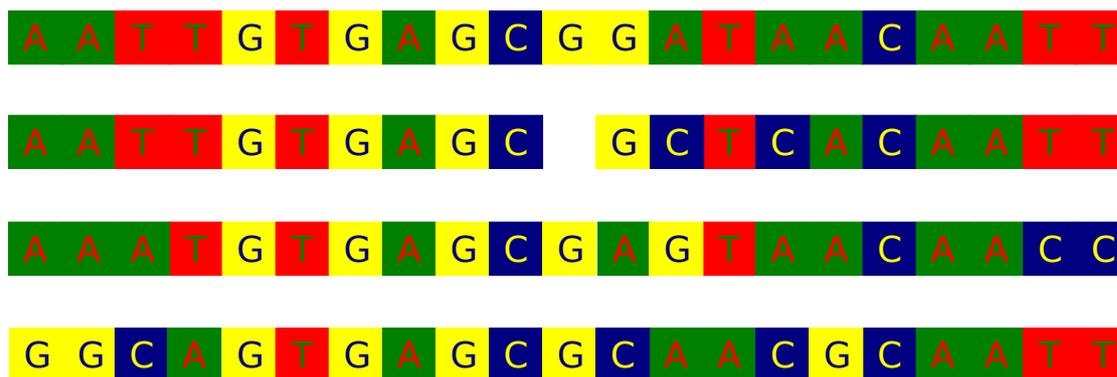


Figure 1.4: Nucleotide sequences of the three natural operators $O1$ (first row), $O2$ (third row) and $O3$ (last row) to which the lac repressor binds specifically. In the second row the sequence of the artificial operator, O_{sym} , is given. Note that in the case of $O2$ the sequence of the "lower strand" is given for better comparability and that the sequence of the symmetric operator is one nucleotide shorter than the naturally occurring ones.

There is a sequence which binds it even stronger and in the *E. coli* genome there are two auxiliary operators.

1.4.1 The symmetric operator O_{sym}

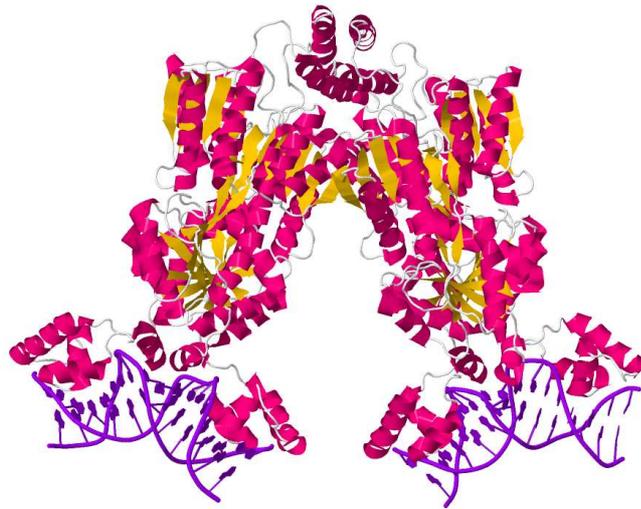
In 1983 an artificial nucleotide sequence was constructed to which the lac repressor binds even stronger than to the naturally occurring $O1$ [42]. The nucleotide sequence of this operator, which is commonly denoted as O_{sym} is given in the second row of Fig. 1.4. Note that it is one nucleotide shorter than the naturally occurring sequence of $O1$ which is shown in the first row of this figure.

The reason for its name is its high degree of symmetry: it is an inverted repeat of the left half of $O1$. This means that its right half is obtained by reflecting it with respect to the middle and by simultaneously replacing each nucleotide with its complementary one. For a moment we take it for granted that the repressor is able to detect the nucleotide sequence in some way. How this is possible will be detailed in section 1.5.

1.4.2 The auxiliary operators

Apart from the main operator, there are two so-called *auxiliary operators*, which will play an important role in chapter 5. Their nucleotide sequences are given in the two last rows of Fig. 1.4. A quick inspection of this figure already tells that $O2$ (third row) shares more nucleotides with the main operator $O1$ (first row) than $O3$ (last row) does. This observation will be quantified in chapter 5.

The stronger auxiliary operator, which we nowadays know as $O2$, was discovered in 1974 by William S. Reznikoff, Robert B. Winter and Carolyn K. Hurley [43]. They found that the affinity of the repressor for this site is approximately 1/30 of the affinity for the main operator and that it lies within the *lacZ* gene. Erroneously, they assumed that its



Jmol

Figure 1.5: The LacI tetramer bound to DNA based on the structure obtained by the group of Klaus Schulten and deposited in the PDB with ID 1Z04 [22, 45]. The image was created with the software Jmol.

rather high affinity can be explained by a single base pair mutation, but they deduced correctly that it is improbable that such a strong binding site is there as though by chance. However, they were unable to pin down its exact role and accordingly in the following *O2* and the even weaker *O3* were called pseudo-operators in a rather deprecatory manner.

Only 16 years after *O2* was found, Benno Müller-Hill and co-workers finished a paper with the remark that a more appropriate name for them would be auxiliary operators [44]. The full quotation will appear below, but what had happened in between? In order to understand this, we first have to study the structure of the main actor in this study, the lac repressor.

1.4.3 Structure of the lac repressor

In its natural form, the lac repressor is a tetramer, sometimes also called a dimer of dimers [45]. Early on, it was noted that a single dimer, i.e. two subunits, has the ability to bind to DNA [46]. In Fig. 1.5 the lac repressor can be seen while it is bound to DNA and assumes a "V"-shaped form.

Both polypeptide arms of the "V" are held together by a four-helix bundle domain [47]. Apart from that, each arm accommodates a core and a head group which is able to bind to DNA [45]. Importantly, the core contains a binding site, called lactose-binding pocket [45]. If lactose binds there, the stability of binding to DNA is reduced, such that the repressor dissociates and the operon is induced [45]. In general, the structure is rather flexible, allowing the repressor to search for a second operator while it is bound to the first one. This explains why the structure of the lac repressor enables it to bind to two stretches of DNA at the same time, folding the intervening DNA into a loop.

However, it is a priori unclear what the biological function of such loops is. This will be discussed in the following subsection.

1.4.4 Looping

First of all, it has to be noticed that there are many types of DNA loops. They can be formed by two proteins which bind to different binding sites and subsequently to each other or—as is the case for the lac repressor—they can be due to a single protein with two binding patches [48]. In principle, there can be various biological reasons why DNA looping mediated by a DNABP is advantageous, see e.g. the review article by Robert Schleif [48].

The most important one for our purposes is that it increases the local concentration of a DNABP¹⁰. This can be understood with the following observation: if a protein can bind to two binding sites which are n base pairs apart, binding to one of them guarantees that the maximal distance between the protein and the other binding site is n base pairs. This distance is often shorter than a typical distance within the cell resulting in an increase of the effective concentration. Thus, binding sites can be saturated at lower protein concentrations than the ones which would be needed with proteins lacking the ability to form loops [48]. This is particularly important, since if all proteins that need to perform tasks in a living cell were produced at high rates, this would inevitably lead to jamming effects. These effects build a whole branch of biophysics under the name of *macromolecular crowding*, see e.g. the reviews [50–52]. The impacts of crowding on our model will be mainly discussed in chapter 4.

Looping in the lac operon Based on his earlier findings that dimeric lac repressors can bind to DNA, in 1977 Jürgen Kania proposed in a joint work with Benno Müller-Hill that the lac repressor in its tetrameric form is able to bind to two stretches of DNA simultaneously [53]. Building on this observation, the most thorough experiment studying the role of each of the three operators appeared in 1990. Therein, Benno Müller-Hill and co-workers used eight plasmids *in vivo*, in which the repression due to all combinations of active or inactive operators was studied [44]. Operators were inactivated by site directed mutagenesis. They compared the expression of β -galactosidase under induced conditions (i.e., in the presence of 1 mM IPTG) and under repressed conditions (without inducer) and calculated their ratio.

While the intact operator region was able to repress the expression by a factor of 1300, deleting one of the two auxiliary operators only mildly decreased the repression to a factor of 700 or 440. However, deleting both auxiliary operators reduced this value enormously to 18. Additionally, whenever *O1* was deleted there was little to no repression potential left [44].

They interpreted the strong repression when *O1* and at least one auxiliary operator is present as resulting from a configuration in which the lac repressor binds to two operators simultaneously. This hypothesis was further corroborated by repeating the experiment with dimeric repressors which are not able to form tetramers. Underlining their expectation, repression by LacI dimers in the presence of *O1* and independent of

¹⁰For this effect several denominations exist, for example cross-talk, cooperativity or recruitment. See the discussion in [49].

the presence of O_2 and O_3 was comparable to the repression by tetrameric LacI when only O_1 was present [44].

Oehler and co-workers closed their paper with a few speculations concerning the evolution of the lac operon: they motivate the observation that on the one hand neither O_1 nor one of the auxiliary operators evolved to O_{sym} and that on the other hand repressors need to be tetramers to tap their full repression potential [44]:

”Here, as elsewhere, evolution rather than favouring the perfection of a simple system (here the dimeric Lac repressor and the dyadic symmetric operator) has instead favoured a cooperative system (here tetrameric Lac repressor and three lac operators). The ‘pseudo-operators’ betray their name and should be called auxiliary operators.“

Up to now, we tacitly assumed that the lac repressor is able to read out the nucleotide sequence of DNA. But how is it made sure that a certain functional sequence only appears once in a bacterial genome?

1.4.5 Connection to information theory

That finding a unique binding site in a genome can also be considered from an information theoretic point of view was already recognised by Walter Gilbert and Benno Müller-Hill in 1967. Without detailing their calculation they stated that in order to select a unique binding site in *E. coli*, which was back then thought to have a genome consisting of 3×10^6 base pairs, a protein must recognise approximately a dozen bases [36]. This can be rationalised by noting that there are $4^{11} \approx 4.2 \times 10^6$ ways to write a word consisting of eleven letters with an alphabet of four letters, and $4^{12} \approx 1.7 \times 10^7$ ways for a twelve letter word.

This calculation, however, relies on the assumption that the occurrence of bases is completely random and that every base is perfectly recognised. As Gilbert and Müller-Hill noticed, if the second assumption is not true, the recognition region must be larger [36]. Related questions will be discussed in more detail in chapter 5.

1.5 Sequence specificity and non-specific binding

Above it was stated that the lac repressor is able to bind *specifically* to certain nucleotide sequences. When comparing the sequences of the three natural operators with the symmetric operator, we already noted that even in the presence of some deviations from the ”perfect“ binding motif, the lac repressor can still be able to bind tightly to DNA. But before we describe how DNABPs are able to interact and to detect specific nucleotide sequences, we have to answer one fundamental question: what happens if the underlying sequence is completely different? Will the repressor still be able to bind to DNA or will it completely lose its affinity?

It was already recognised by Arthur D. Riggs and co-workers in an article which is central to this study that the lac repressor has a general affinity for DNA¹¹ [55]. This

¹¹Even before that, it was David Pettijohn and Tomoya Kamiya who showed that RNAP can bind non-specifically to DNA and that the affinity depends on the ionic strength [54].

general affinity is usually referred to as non-specific binding and describes the situation when a DNABP cannot only bind to its specific target sequence, but also to other stretches of DNA. Riggs et al. noticed that long-ranged electrostatic forces dragged the protein of their study towards DNA resulting in a rather weak non-specific affinity. Besides, they interpreted specific interaction with a sequence as resulting from "reading" the edges of the corresponding bases in the minor and major groove, assuming that the four possible base pairs differ enough in these edges to be distinguished.

The continuous transition from specific to non-specific binding was concisely described by Peter H. von Hippel and Otto G. Berg [56]:

"There is some finite level of affinity of the protein for the 'correct' site and some lower (but non-zero) and progressively decreasing affinity for other sites with decreasing degrees of homology with the correct one."

This is on the whole still the current opinion in the field: non-specific binding is mostly mediated by electrostatic forces. Conversely, for specific binding it is exploited that the double helix' outer part is "studded with sequence information" [13]. This was affirmed by nuclear magnetic resonance (NMR) studies: comparing their chemical shift perturbations and the broadening of lines, amino acids could be grouped two-fold: some contact the DNA mostly with their side chains (hydrophobic interactions and water-mediated hydrogen bonds) while others build direct hydrogen bonds with the backbone [57]. The specific binding energy is the superposition of many weak contacts, e.g. hydrogen bonds, ionic bonds or hydrophobic interactions, which only together yield strong binding energies [13]. Importantly, the underlying sequences can be distinguished without opening the helix as it was the case when the sequence is actually transcribed by the RNAP [13]. Von Hippel and Berg also estimated the specific and non-specific binding energy to be ≈ 17 kcal/mol and ≈ 7 kcal/mol in the physiologically relevant regime [56]. Interestingly, this estimate is in the ballpark of earlier guesses by Gilbert and Müller-Hill [36].

An important point for non-specific interaction is that due to its electrostatic nature, and more exactly since it is mainly mediated by the release of many counter ions from DNA, the binding affinity heavily depends on the ionic strength of the environment [58]. High salt concentrations imply weak non-specific binding and low salt concentrations strong non-specific binding. However, not only the salt concentration but also the type of salt involved matters. For example, Mg^{2+} ions bind more tightly to DNA than monovalent ions [58].

From early on, it was discussed if non-specific and specific binding occurs in different binding modes. For example, Winter and co-workers speculated that in the absence of a target sequence DNABPs maximize the interaction with the DNA backbone, while in the presence of it these electrostatic interaction are reduced to have more direct interaction with the nucleotides [59]. This will be discussed in more detail in section 2.4.4.

The significance of the occurrence of non-specific binding was in particular noted by von Hippel and co-workers who stated that any quantitative model of repression has to take non-specific binding into account [60]. At first sight, the huge difference in binding affinity for specific and non-specific sites seems to imply that non-specific binding can be ignored. However, when every base pair of the circular *E. coli* genome is considered as the leftmost position for non-specific binding, it becomes obvious that there are nearly

five million possible non-specific binding position, which compete with three specific ones. This calls for a description in terms of a thermodynamic model which will be delivered in the final subsection of this chapter.

1.5.1 Thermodynamic models

All thermodynamic models aiming at a description of the lac repressor system rely on the assumption that the cell is in equilibrium. Then the expression output of the lac operon is directly proportional to the fraction of time the main operator remains unoccupied by the repressor which in turn depends on the concentrations of binding sites and repressors in the cell. This was noted already by Gilbert and Müller-Hill in 1967 [36].

We now follow the description presented in 1986 by Peter H. von Hippel and Otto G. Berg who denoted the fractional saturation of the binding site by θ_s and consequently the selection factor by $x = \theta_s/(1 - \theta_s)$ [56]. Here, in principle θ_s can attain values between 0 (low saturation) and 1 (infinitely high saturation). Experimentally, it is known that a fully induced system produces approximately thousand times more proteins of the lac operon as compared to a repressed one. This implies $x \approx 1000$. Denoting the total repressor concentration by R_T and by D_i the concentration of binding sites characterised by a binding constant K_i , one obtains [56]:

$$R_T = R_F + \sum_i \frac{K_i D_i R_F}{1 + K_i R_F}, \quad (1.1)$$

where R_F denotes the concentration of unbound repressors and where the sum over i comprises non-specific and specific binding sites. With the above mentioned selection factor x , this can be rewritten as [56]:

$$R_T = \frac{x}{K_s} + \frac{x D_s}{1 + x} + \sum_{\text{ps}} \frac{x D_i}{x + K_s/K_i} + \sum_{\text{ns}} \frac{x D_i}{x + K_s/K_i}, \quad (1.2)$$

where K_s refers to the specific binding constant. Again, the first term corresponds to unbound repressors. However, in this formulation the second term on the right hand side of Eq. (1.1) was split up into three terms on the right hand side of Eq. (1.2). The first one of those describes the TFs which are bound to the main operator, the middle one those which are bound to pseudo-sites which have rather strong binding energies and the last one the ones bound to non-specific sites. Using that for pseudo-sites we have $K_s/K_i \ll x$, while for truly non-specific sites $x \ll K_s/K_i$ and that $x \gg 1$ and therefore $x/(x + 1) \approx 1$, yields

$$x \approx \frac{K_s}{1 + \sum_{\text{ns}} K_i D_i} [R_T - D_s - D_{\text{ps}}], \quad (1.3)$$

where D_{ps} is a short for the sum of concentrations of pseudo-sites: $D_{\text{ps}} = \sum_{\text{ps}} D_i$.

This equation has a straightforward interpretation for the expression output: the selection factor, x , is simply the product of an effective repressor concentration with an effective specific binding constant. Thus, a high abundance of non-specific binding sites can greatly reduce the effective binding constant, while strong pseudo-sites reduce the amount of repressors which are free to find the main operator. Note that in 1986 when

this article was published the role of the auxiliary operators was not yet fully resolved. More recent thermodynamic models take the looped states into account, see e.g. [61] and references therein.

In general, thermodynamic models are widely applied to describe cellular systems and often highlight the importance of non-specific binding, see e.g. [61–64] and references therein. Even if they are able to describe the physical situation reasonably well, one has to keep in mind that they implicitly assume that the cell is in equilibrium which is a questionable assumption. Furthermore, the "occupancy hypothesis" on which all thermodynamic models of gene regulation rely has very recently been called into question [65].

In the following "transition" chapter we describe how the binding of a DNABP to a target sequence can be considered as a biological search process, before chapters 3, 4 and 5 describe the main findings obtained during this PhD project.

2 Biological search processes

In the previous chapter this work focused mainly on biological issues. It was emphasised that for the survival of a bacterial cell, it is utterly important to regulate gene expression quickly and reliably in reaction to some signal. An important ingredient is the binding of a TF to (one of) its target sequence(s) on DNA. But what can a theoretical physicist contribute to this topic? The answer is that from a technical point of view, a TF searching for a target on the DNA is nothing but a specific realisation of a search process. Of course, there is more to gene expression and regulation than just the action of TFs, but efficient search of TFs is a pre-requisite for it. Besides, the techniques we employ here can be applied to other search processes as well.

In 1967 Walter Gilbert and Benno Müller-Hill published an article entitled “The lac operator is DNA” in which they estimated—without detailing the calculation—that the association of lac repressor to its operator is diffusion-limited and that it will occur at a maximal rate of $10^8 \text{ M}^{-1}\text{s}^{-1}$ [36]. The term *diffusion limit* can be found basically in any publication concerning this topic.

In the following section we clarify what is meant by this and why it is highly disputed even nowadays (compare also subsection 2.5.2). Section 2.2 will present early theoretical descriptions of this search process, while section 2.3 is devoted to the so-called facilitated diffusion model introduced in 1981 by Berg, Winter and von Hippel. Then, section 2.4 will summarise experimental and theoretical approaches of the last decade and finally, section 2.5 will describe which features all variations of the facilitated diffusion model have in common and what is criticised about it.

2.1 The diffusion limit

To understand how Gilbert and Müller-Hill obtained their estimate, we have to travel back in time even further. In a seminal work published nearly a century ago, the Polish physicist Marian Smoluchowski studied a bimolecular reaction in a system in which a substance characterised by a diffusion coefficient D_3 fills the infinite space with a homogeneous concentration [66]. If at time $t = 0$ a perfectly absorbing sphere of radius a centred around the origin is introduced to the system, the rate at which the substance diffuses to the sphere is simply given by¹:

$$k_{\text{Smol}} = 4\pi D_3 a. \quad (2.1)$$

In terms of the biological search vocabulary this is the product of the reaction radius of the target (a) the diffusion coefficient of the searcher (D_3) and the full solid angle (4π).

¹Note that this rate has the physical dimension m^3/s , that is volume per time unit and is usually given by experimentalists in units of $\text{M}^{-1}\text{s}^{-1}$ (see above). Here 1 M (= “1 molar”) refers to a concentration of one mole per litre. Besides, here and throughout this work the notation in formulae has been adjusted to the one chosen in our publications.

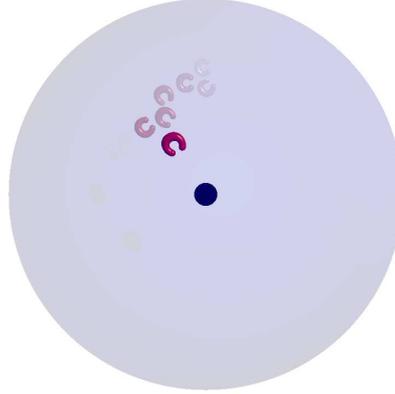


Figure 2.1: Scheme of a search process in which a TF, represented by the U-shaped particle, is searching for a spherical target in a spherical reaction volume.

To be more precise, D_3 denotes the sum of the diffusion coefficients of searcher and the one of the target. However, in the context of protein-operator associations, usually the diffusivity of the target is neglected, since it is part of a huge DNA molecule.

The situation is schematically depicted in Fig. 2.1 where the U-shaped TF searches for the spherical target in a reaction volume which is spherical, too. Herein, the partially transparent shapes represent previous positions of the TF. Before stepping further it is important to note that this result describes a rather idealized situation. In a real biological system, there are further influences: most importantly, both searcher and target may have charges leading to a repulsive or attractive electrostatic interaction. Besides, the surfaces of both particles which are in general not spherical may not be uniformly reactive [67]. The corrections to Smoluchowski's result which these factors imply will be discussed in subsection 2.1.3. Before we detail this, we study how the result can be generalised to finite systems using the so-called first passage time formalism.

2.1.1 First passage time formalism

In 1980 Attila Szabo together with Klaus and Zan Schulten discussed in terms of the first passage time formalism how to calculate the mean time τ that a particle needs to locate a target when it diffuses with a possibly space-dependent coefficient $D(\mathbf{r})$, and experiences a potential U at the same time [68]. For a particle starting at \mathbf{r}_0 , this amounts to solving the following equation for the mean first passage time τ :

$$L^+(\mathbf{r}_0)\tau(\mathbf{r}_0) = -1, \quad (2.2)$$

which involves the adjoint operator defined via $L^+(\mathbf{r}) = \nabla \cdot D(\mathbf{r})\nabla - \beta D(\mathbf{r})(\nabla U) \cdot \nabla$. Now we consider a system with spherical symmetry in d dimensions, with a reflecting boundary at a radial distance of R [68]:

$$\left. \frac{d\tau(r_0)}{dr_0} \right|_{r_0=R} = 0. \quad (2.3)$$

Furthermore, at the target (radial distance a), a radiative boundary condition characterised by the reaction rate, κ , is assumed [68]:

$$\left. \frac{d\tau(r_0)}{dr_0} \right|_{r_0=a} = \frac{\kappa\tau(a)}{D(a)}. \quad (2.4)$$

Taking as initial condition the equilibrium distribution in the absence of a target,

$$p_{\text{eq}}(r) = \frac{r^{d-1} \exp(-\beta U(r))}{R \int_a^R dy y^{d-1} \exp(-\beta U(y))}, \quad (2.5)$$

one obtains the following general result for the mean first passage time [68]:

$$\tau = \int_a^R dx [D(x)p_{\text{eq}}(x)]^{-1} \left[\int_x^R dy p_{\text{eq}}(y) \right]^2 + [\kappa p_{\text{eq}}(a)]^{-1}. \quad (2.6)$$

In the case of free diffusion in 3D, $U = 0$ and $D = D_3$, and using the dimensionless distance $x = a/R$, one obtains [68]:

$$\frac{\tau D_3}{R^2} = (1-x)^2 \frac{5+6x+3x^2+x^3}{15x(1+x+x^2)} + D_3 \frac{1-x^3}{3\kappa R x^2}. \quad (2.7)$$

Since the system volume, V , can be written as $V = (1-x^3)4\pi R^3/3$, in the limit of a fast intrinsic reaction rate ($\kappa \rightarrow \infty$) the association rate $k_a = V/\tau$ results as²

$$k_a = 4\pi D_3 a \frac{5(1-x^3)(1+x+x^2)}{(1-x^2)(5+6x+3x^2+x^3)} \xrightarrow{x \rightarrow 0} 4\pi D_3 a, \quad (2.8)$$

where the last limit for infinite system size underlines that this result is equivalent to Smoluchowski's formula with all the terms involving x accounting for finite size effects.

Since it will be of use later on, we also give the result for free diffusion in the 1D interval from a , where the absorbing boundary lies, to R , where the boundary is reflective. One obtains using $U = 0$ and $D = D_1$ in Eq. (2.6) [68]:

$$\frac{\tau D_1}{R^2} = \frac{(1-x)^2}{3}. \quad (2.9)$$

In the limit $a \rightarrow 0$, i.e. $x \rightarrow 0$, and $R = L$, this takes the more common form:

$$\tau = \frac{L^2}{3D_1}. \quad (2.10)$$

²A similar formula with appropriately adapted boundary conditions will be encountered in chapter 4.

2.1.2 Experimental results by Riggs et al.

In 1970, Arthur D. Riggs and co-workers studied the association of lac repressor with its operator in an *in vitro* experiment. Using radioactively labelled DNA from a phage called $\lambda\phi 80d$ which contains the whole natural *lac* region, they measured an association rate of $7 \times 10^9 \text{ M}^{-1}\text{s}^{-1}$ [55]. Conversely, they compared this value to a theoretical estimate using Smoluchowski's result, Eq. (2.1). Plugging in the values $a = 0.5 \text{ nm}$ and $D_3 = 50 (\mu\text{m})^2\text{s}^{-1}$, they obtained $k_{\text{Smol}} \approx 10^8 \text{ M}^{-1}\text{s}^{-1}$ in accordance with Gilbert and Müller-Hill's previous estimate [36, 55].

This result came as a surprise, since the reaction apparently proceeds nearly two orders of magnitude faster than theoretically expected. Of course, the authors tried to explain this discrepancy. They saw the reason for this remarkable acceleration in an electrostatic interaction between the phosphate groups in the operator DNA and some positively charged site on the repressor [55]. Even though they did not have an extension of Smoluchowski's formula for charged macromolecules at hand, they supposed that it would result in a substantial acceleration of the association. How Smoluchowski's formula changes when electrostatic interaction between the searching particle and the target is taken into account will be discussed in subsection 2.1.3.

Interestingly, Riggs et al. also discussed the acceleration in terms of an "extreme model of oriented diffusion" whose basic concept is very similar to the facilitated diffusion model which was introduced several years later [55]. According to this model rather than diffusing in 3D as described by Smoluchowski's theory, the TF is drawn towards the DNA by electrostatic forces and subsequently hops or rolls along DNA. This reduces the dimensionality of the search process, an effect that will be discussed more generally in subsection 2.2.1.

However, Riggs and co-workers dismissed this model for several reasons. They deemed non-specific binding to be too weak and thought that the direct electrostatic interaction speeds up the reaction sufficiently, such that an additional acceleration via the oriented diffusion model was unnecessary³. The most interesting counter-argument they brought up was that the rolling mechanism implies a dependence of the association rate on the length of the flanking DNA which they did not detect in their earlier results, where DNA segments with a molecular weight down to 10^6 were studied⁴ [55, 70]. At the same time they proposed to check this with even shorter DNA segments, which was later done by Reimund Fickert and Benno Müller-Hill [71].

2.1.3 Extension of Smoluchowski's formula

As already mentioned, the original result of Smoluchowski was obtained for rather idealised conditions. To describe more realistic situations, the fact that both searcher and target may be charged and can have a non-uniform reactivity is often taken into account phenomenologically by multiplying the right hand side of Eq. (2.1) with an electrostatic and a steric factor [67, 72]. As appealing as it may seem to wrap up the steric and

³Interestingly, only two years later Lin and Riggs mentioned this model again to explain the rapidity of the specific association [69].

⁴Suzanne Bourgeois and co-workers mentioned preliminary results obtained with a DNA of molecular weight of less than 10^5 , which still showed operator activity [70].

electrostatic influence into simple multiplicative terms, it has to be said that a concrete calculation appears intractable.

In the case of the steric factor, one has to note that collisions between macromolecules are not elastic [67]. Thus, after a first contact there may be several repeated collisions such that one can assume that a non-uniform reactivity is not decisive. Conversely, much of the modern criticism concerning the diffusion limit comes from the fact that this formula neglects possible charges of the particles (compare also subsection 2.5.2) [72, 73]. In fact, the non-specific interaction between the repressor and the DNA backbone is mostly electrostatic. Such electrostatic effects can be included into Smoluchowski theory. With an attractive interaction potential $U(r)$ in 3D, one obtains the following generalisation of Smoluchowski's result [74]:

$$k_{\text{Smol}} = 4\pi D_3 a_{\text{eff}}, \quad (2.11)$$

where a_{eff} denotes the effective range of the potential for a particle starting at a radial distance r_1 and with an absorbing boundary at distance a [74]:

$$a_{\text{eff}} = \int_a^{r_1} \frac{\exp U(r)}{r^2} dr. \quad (2.12)$$

Thus, an interaction potential increases the effective size of the target which is perceived by the searching particle. Note that Eq. (2.12) returns the classical result in the absence of a potential, $U(r) = 0$, when choosing an initial condition which fulfills $r_1 \gg a$.

However, the interior of a cell is definitely not a vacuum. Therefore one cannot directly use this result. In fact, the electrostatic interaction is (partially) neutralised by the presence of counter ions. Estimates tell that at a salt concentration of 0.1 M interactions are screened at distances of a few nm [75,76]. Therefore it seems inappropriate to assume that direct long-range electrostatic interaction between the target and the searching particle speeds up the search process considerably.

2.2 Early studies

The surprise of the high association rate measured by Riggs and co-workers soon ignited the interest of theoretical physicists. Two researchers, Peter H. Richter and Otto G. Berg, and their co-workers had an outstanding role in developing these early studies. While Richter and co-workers used a steady-state approach, the description of Berg and co-workers was based on a calculation of mean association times. From a technical point of view, both methods are equivalent and rely on the existence of a dominant relaxation time [77]. Importantly, both are based on the notion that motion in reduced dimensionality can be beneficial for the search process. Thus, before we present their approaches in subsections 2.2.2 and 2.2.3, we first study this concept in general terms.

2.2.1 Reduction of dimensionality

That using the strategy to search for a target in reduced dimensionality can be advantageous for target detection was discussed decades before experiments with the lac repressor

were performed. In 1921 the Hungarian mathematician George Pólya showed that there is a fundamental difference between random walks on lattices of different dimensionality: while in 1D and 2D a particle will return with certainty to the point where it started its random walk, in more than two dimensions this probability is less than unity⁵ [78].

In other words, a random walker on a line will often return to its starting point. This redundancy can be made palpable by observing that for such a walker the probability to move closer to the target is exactly as high as the one for moving further apart. Pólya derived probabilities of return to the starting point. The explicit rates were considered by Adam and Delbrück two years before the experimental results of Riggs et al. were published [80].

They found that in d dimensions the mean time of diffusion to a target, $\tau^{(d)}$, can be written as the product of a distance factor and a tracking factor, which is a function of the ratio of the linear size of the system, b , and the target size a [80]:

$$\tau^{(d)} = \frac{b^2}{D_d} \cdot f^{(d)}\left(\frac{b}{a}\right). \quad (2.13)$$

As usual D_d denotes the diffusion coefficient in d dimensions. This result is equivalent to the ones obtained with the FPT approach in the previous section.

Based on these observations, they compared diffusional association times in two systems. The first one consisted of two concentric spheres of radii a and $b > a$, where the smaller sphere represents the target and the larger sphere the outer boundary of the system. The second system consists of the same larger sphere of radius b , but now a membrane lies in its equatorial plane. In the membrane's centre resides the target which again has size a (compare Fig. 2.2). In this case, the search has two phases: first the particle has to bind to the membrane and then it diffuses in 2D on this surface until the target is found.

The ratio between the diffusion times in both cases depends solely on the ratio of the diffusion coefficients in 2D and 3D and on the relative size of the target with respect to the system size. When diffusion on the membrane is not too slow and when the target is rather small, they found that “combined space and surface diffusion is favorable” [80]. Importantly, this strategy is only advantageous in a certain parameter regime and as evidenced by their calculations the search can even be slowed down in other regimes. Furthermore they detailed that the strategy might be particularly important “in the case of [...] *small* amounts of [...] regulators” [80]. Differences in strategy for biological particles which are present in different concentrations to find a target as soon as possible were indeed recently observed experimentally [81]. As a side remark we note that the work of Adam and Delbrück was the first on this topic where the term “antenna” was mentioned.

2.2.2 The contributions of Peter H. Richter

In 1974, Peter H. Richter and Manfred Eigen readopted the model that Riggs and co-workers had rejected in their seminal paper [82]. They were not content with Riggs

⁵For example, in the relevant case of a three-dimensional cubic lattice, this probability amounts to ≈ 0.3405 .

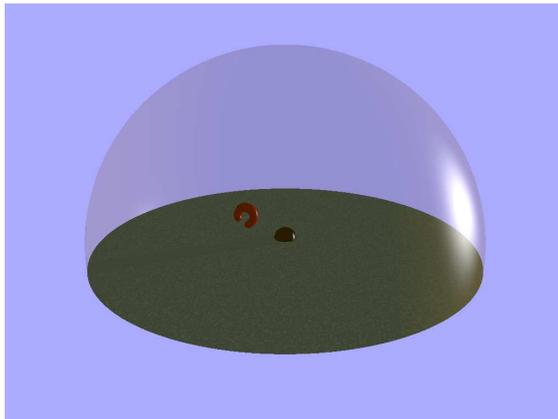


Figure 2.2: Scheme of a search process involving motion in reduced dimensionality. The target resides in the centre of a membrane which in turn lies in the equatorial plane of the spherical search volume.

et al.'s explanation that the direct electrostatic interaction is the main factor for the high association rate, and expected that it should play a more prominent role in the inverse process when a particle dissociates from the target. Based on their opinion that electrostatics mostly influences the non-specific association, they extended the classical Smoluchowski ansatz to spheroidal symmetry.

They considered that the repressors diffuse in 3D and are able to bind non-specifically to DNA. While bound non-specifically the repressor diffuses along the DNA with the 1D diffusion coefficient⁶ D_1 , but dissociation back to the solution occurs at a rate of k_{off} . The acceleration of the association reaction is explained by the fact that a typical distance which is travelled during the bound phase is approximately⁷ $\ell = \sqrt{D_1/k_{\text{off}}}$. Thus, since the operator can be reached via 1D diffusion from both sides, its effective size is proportional to twice this value. Because a cylinder can be approximated by a long prolate spheroid, the relevant result for the association rate to a cylinder of length 2ℓ and radius B reads [82]:

$$k_a = 4\pi D_3 \frac{-l_b}{\exp\left(-\frac{l_b}{\ell} \ln\left(\frac{2\ell}{B}\right)\right) - 1}, \quad (2.14)$$

where l_b denotes Bjerrum's length $l_b = -q_D q_R / (4\pi\epsilon k_B T)$ with the charge of the operator q_D , the one of the repressor q_R and the dielectric constant ϵ . This length denotes the distance at which the electrostatic interaction between two charges is $k_B T$. For small electrostatic forces, $|l_b| \ll \ell / \ln(2\ell/B)$, an expansion of the exponential function yields:

$$k_a \simeq 4\pi D_3 \frac{\ell}{\ln(2\ell/B)} \left(1 + \frac{b}{2\ell} \ln\left(\frac{2\ell}{B}\right)\right). \quad (2.15)$$

What is immediately apparent is that the form of this is similar to Smoluchowski's classical result, namely $4\pi D_3$ times an effective target size. In this case, the latter is

⁶Thus, unlike in Adam and Delbrück's original work, now the second search phase does not proceed on a two-dimensional membrane, but along the one-dimensional contour of the DNA chain.

⁷We note that some authors include a factor of two within the square root in the definition of ℓ .

represented by a slightly intricate expression. But obviously for appropriately chosen values of k_{off} and D_1 , ℓ can be much larger than the “naked” target size. Thus, the non-specific DNA next to the target acts as an “antenna” which collects incoming particles.

Importantly, Richter and Eigen were able to explain the strong dependence of the specific association rate on ionic strength. Due to the experimental studies of Lin and Riggs [69], it was known that a small increase in ionic strength causes a large increase of the unspecific dissociation constant. This has an impact on the specific association rate, since larger dissociation rates reduce the typical distance, ℓ , which is travelled along DNA. This results in a smaller effective target size. Furthermore, Richter and Eigen rejected the final argument of Riggs and co-workers by stating that the short DNA segments that they mentioned as a testimony against a length dependence were simply not short enough to see such a dependence [82].

Four years later, together with Rudi Schraner, Peter H. Richter published a refined model, termed “guided diffusion” [83]. In this version, they accounted for the criticism uttered by Otto G. Berg and Clas Blomberg that the original model of Richter and Eigen did not explicitly couple one-dimensional diffusion along DNA to the bulk diffusion [82, 84]. Introducing the radius R of an outer cylinder which fills the reaction space and where the concentration of searching particles is fixed, they obtained a result which was again similar to Smoluchowski’s:

$$k_a = 4\pi D_3 \frac{\ell \tanh(L/\ell)}{\ln(R/B)}, \quad (2.16)$$

where $2L$ denotes the length of the cylinder and enters the last term which describes the effective target size. Furthermore, applying the usual notation, $\beta = 1/(k_B T)$, Schraner and Richter considered that Eq. (2.16) has to be multiplied by $-\beta U(b)/(1 - \exp[\beta U(b)])$ to take into account electrostatic effects explicitly. Alternatively, Eq. (2.16) is multiplied by $(1 + 1/[kb \ln(R/b)])^{-1}$ to consider a non-specific association which is not exclusively diffusion-limited, but characterised by the intrinsic association rate k .

Interestingly, Schraner and Richter were the first to mention a fact which is similar to what is nowadays known as the *speed-stability paradox* (compare section 2.4.4). Namely, that it might be impossible to have a very large one-dimensional range, $\ell = \sqrt{D_1/k_{\text{off}}}$, since the requirements of large non-specific affinity (low values of k_{off}) and large mobility (large values of D_1) are somewhat opposite [83]. Besides, they argued that both for very long and very short DNA segments, their result must be refined. In the first case, DNA will no longer be adequately modelled as a cylinder, but rather as a random coil and in the second case, the diffusion constant of the DNA cannot be neglected (compare the discussion following Eq. (2.1).)

2.2.3 The contributions of Otto G. Berg

Responding to Peter H. Richter’s first paper published two years earlier, in 1976 Otto G. Berg and Clas Blomberg studied the model in more detail [84]. This was actually the first one in a whole series of papers, culminating in three papers by Robert B. Winter, Peter H. von Hippel and Otto G. Berg published in a single issue of the journal *Biochemistry* in 1981 [59, 75, 85]. The latter described the *facilitated diffusion model* which is the foundation of nearly all more recent works. It will be discussed in section 2.3.

While Berg and Blomberg agreed with the general concept of the steady-state model proposed by Richter and Eigen, they criticized that two points were neglected which slow down the search. First, the DNA far away from the target traps the repressor during the search process, and second, even after landing near the target, some time will pass while the TF moves towards the target.

Berg and Blomberg took these effects into account and made a few simplifying assumptions concerning the arrangement of DNA. Namely, they used a closed-cell approach, in which several straight DNA segments of length $2L$ and radius b are aligned in parallel and regularly spaced [84]. Around each DNA chain an imaginary cylinder is drawn with radius R , such that the whole reaction volume is filled. The TF starts at a random position in the bulk solution and searches for the target, which resides at the centre of the DNA chain. At the outer cylinder there is a reflective boundary condition, motivated by the observation that in such a parallel arrangement statistically for every particle which attempts to move to a neighbouring cylinder volume, there will be another one which will move in exactly the opposite direction⁸. Within this model the mean time of target association, τ , can be written as:

$$\tau = \tau_1 + N(\tau_2 + \tau_3). \quad (2.17)$$

These terms have a straightforward interpretation: τ_1 is the mean time that passes until the TF associates for the first time non-specifically with the DNA. After that there are N consecutive rounds in which on average the particle diffuses in 3D for a time span $\tau_2 = 1/(4\pi D_3 L b k n_0)$ and slides one-dimensionally along DNA for a time span $\tau_3 = 1/k_{\text{off}}$ [86]. Here n_0 denotes the density of operators and the conventional association rate can be obtained from the mean search time via $k_a = 1/(\tau n_0)$.

We do not discuss the individual form of the terms N and τ_1 in detail, since on the one hand we will encounter modified versions of them in chapter 3 and on the other hand below we will give the corresponding terms which appeared in a follow-up paper published one year later. Intriguingly, plugging in the experimental results of Riggs et al., they were able to predict a one-dimensional diffusion coefficient of $0.3 \frac{(\mu\text{m})^2}{\text{s}}$ many years before such values were directly measurable.

In the above mentioned first follow-up paper, Berg and Blomberg considered a system with slightly different boundary conditions [86]. At the outer cylinder instead of the reflective boundary condition, they introduced an absorbing boundary condition. However, particles which reached this radial distance from the DNA were not assumed to be lost, but rather were reintroduced to the system with an arbitrary coordinate along the cylinder. This means that a particle which reaches this distance loses all its memory of where it dissociated before. Technically, the flux of particles which have dissociated, is partitioned into those who return to the chain without ever reaching the outer cylinder boundary, $\phi(t)$, and those who did, $\psi(t)$. These two functions are combinations of modified Bessel functions. We do not state their explicit form, but in the analogous case of our generalised model presented in chapter 3, they will be presented. With these assumptions the Laplace transform of the flux of particles into the target, $\tilde{\Phi}$, is described

⁸In chapter 3 we will use these very boundary conditions in our generalised facilitated diffusion model because they allow for a closed-form solution.

by [86]:

$$\tilde{\Phi}(u) = \frac{L\tilde{G}(u)}{1 + 2 \left\{ u + k_{\text{off}} \left(1 - \tilde{\phi}(u) - \tilde{\psi}(u) \right) \right\} \sum_{n=1}^{\infty} \left(v_1(u, n) + k_{\text{off}} \left[1 - \tilde{\phi}(v_3(u, n)) \right] \right)^{-1}}, \quad (2.18)$$

where $v_1(u, n) = u + \frac{D_1 n^2 \pi^2}{L^2}$ and similarly $v_3(u, n) = u + \frac{D_3 n^2 \pi^2}{L^2}$ depend on the variable u which is the Laplace variable complementary to time t .

Here and throughout this work, tildes denote Laplace transforms whose utility lies in the fact that—if fluxes are considered—their value at $u = 0$ gives the probability with which the target will be detected. Besides, apart from a pre-factor of -1 the value of their derivative at the origin is the (conditional) mean first passage time to the target [84]. In the present case, one obtains $\tilde{\Phi}(u = 0) = 1$, which simply tells that in the finite system with reflective boundaries at the periphery of it, the target will be found with certainty.

Calculating the first derivative of Eq. (2.18) at $u = 0$, one obtains that the mean target association time τ can again be written as in Eq. (2.17), where N has the following form [86]:

$$N = \sum_{n=1}^{\infty} \frac{2k_{\text{off}}}{D_1 n^2 \pi^2 / L^2 + k_{\text{off}} [1 - \tilde{\phi}(D n^2 \pi^2 / L^2)]}. \quad (2.19)$$

Given that $\tilde{\phi}$ has a rather intricate form involving modified Bessel functions, Berg and Blomberg made the assumption that the term $1 - \tilde{\phi}$ can be approximated by $(1 + bk \ln[R/b])^{-1}$, in close analogy to Schraner and Richter's correction term. This corresponds to assuming that the re-association flux to positions which are close to the point of dissociation can be lumped together at the latter [77]. To ease the notation the macroscopic dissociation rate $\Lambda = k_{\text{off}} / (1 + bk \ln[R/b])$ was introduced. Here, macroscopic dissociation is to be understood as opposed to the microscopic one with rate k_{off} , in the sense that Λ/k_{off} describes the fraction of particles which lose correlation with their point of dissociation [77]. In modern terminology this definition allows to distinguish between micro- and macro-hops. One obtains for the mean number of search rounds [77]:

$$N \approx \frac{k_{\text{off}}}{\Lambda} \left[\sqrt{\frac{\Lambda L^2}{D_1}} \coth \left(\sqrt{\frac{\Lambda L^2}{D_1}} \right) - 1 \right]. \quad (2.20)$$

Thus, the number of search rounds which is needed to find the target is mainly determined by the ratio of the half-length of DNA, L , and the effective range of the operator $\sqrt{D_1/\Lambda}$ [77, 82]. For $L \gg \sqrt{D_1/\Lambda}$, one obtains $N \approx k_{\text{off}} \cdot \sqrt{L^2/(\Lambda D_1)}$, and in the opposite limit $\sqrt{D_1/\Lambda} \gg L$ this yields $N \approx (k_{\text{off}} L^2)/(3D_1)$ (compare Eq. (2.10)).

The macroscopic dissociation rate can be generalised to take electrostatic interactions into account by redefining Λ via [77, 87]:

$$\Lambda = k_{\text{off}} \left\{ 1 + bk \exp(-\beta V(\rho)) \int_b^R \frac{1}{\rho} \exp(\beta V(\rho)) d\rho \right\}^{-1}, \quad (2.21)$$

where V denotes the electrostatic potential. For $V = 0$, this yields the previous result, $\Lambda = k_{\text{off}}/(1 + bk \ln(R/b))$.

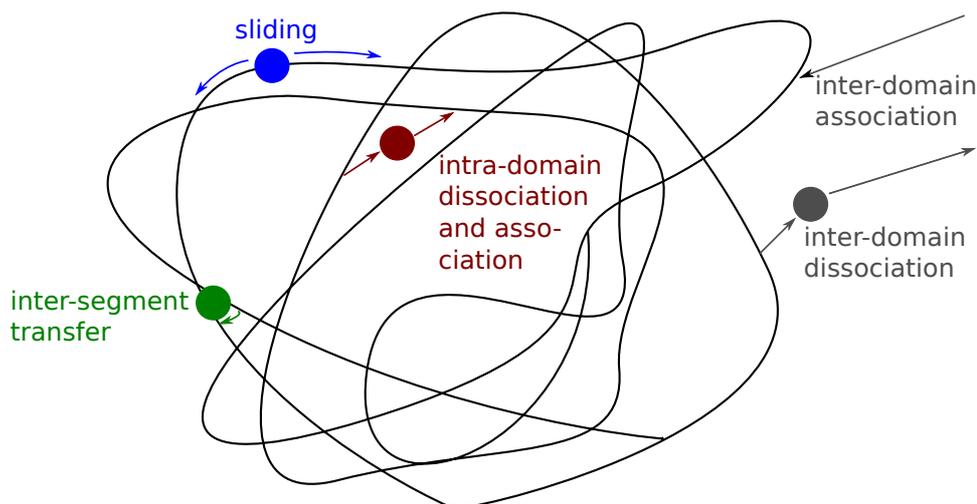


Figure 2.3: Scheme of a searching TF (adapted from [67]). The blue TF slides along DNA. The dark red TF has just dissociated from DNA and since it will re-associate at a nearby position, this constitutes an intra-domain dissociation and association. Conversely, the grey TF leaves the domain of this DNA molecules after dissociation, making an inter-domain dissociation. Finally, the green TF is bound to two stretches of DNA simultaneously, enabling it to perform an inter-segment transfer.

2.3 The Berg-von Hippel or facilitated diffusion model

In a seminal series of papers in 1981 Otto G. Berg, Peter H. von Hippel and Robert B. Winter combined their earlier results into a generalised model and tested its predictive power with experiments [59, 75, 85].

A short note on nomenclature The model that Otto G. Berg and co-workers detailed in this series of papers bears many names. Some refer to it as the Berg-(Winter-) von Hippel model, while most call it the “facilitated diffusion” model. The precursor model introduced by Schraner and Richter was named “guided diffusion”, while Riggs and co-workers spoke of “oriented diffusion” and finally Robijn Bruinsma coined the term “mixed diffusion” [19, 83]. In the following we continue to use the term “facilitated diffusion model”, simply because it is the most common one, even though one might argue that the more neutral term “mixed diffusion” is more appropriate.

Importantly, Berg and co-workers introduced the concept of a “hierarchy of non-specific dissociation rates” in which they distinguished between microscopic, intra-domain and inter-domain dissociations [85].

After having dissociated from DNA, with a rather high probability, the TF will return to a position near the dissociation point. This is a microscopic dissociation, mentioned in the previous section. As implied by the dark red TF in the schematic Fig. 2.3, it can also land on DNA which is close in 3D space, but distant along the contour of DNA. This constitutes an intra-domain dissociation. Finally, inter-domain dissociations

refer to events where the whole domain of a DNA molecule is left (compare the grey TF in Fig. 2.3). Such an inter-domain dissociation is complementary to the association described by τ_1 in Eq. (2.17), which is the mean time of first non-specific association in their previous publications.

In modern terminology, one speaks of “hops” if the re-association point is close to the dissociation point and of “jumps” if these two position are far apart along the contour. Berg and co-workers also introduced the concept of inter-segment transfer (IT). This describes a mechanism, in which a protein that has two binding patches uses them to simultaneously bind to two stretches of DNA which are close in 3D space [85], compare the green TF in Fig. 2.3. If the binding patch which subsequently releases DNA again is the one which was bound in the beginning while the newly built connection remains, the protein effectively translocated along DNA without dissociation. This distinguishes this mode of transport from inter-segmental jumps, in which the TF dissociates and subsequently associates with another stretch of DNA. Within this full facilitated diffusion model, the specific association rate can be written as

$$k_a = \left[(Mk_1)^{-1} + \left(\frac{k_2 D_T K_{RD}}{1 + D_T K_{RD}} \right)^{-1} \right]^{-1}. \quad (2.22)$$

The first term on the right hand side corresponds to the rate for the first non-specific association with DNA, where M denotes the total number of (non-specific) sites and k_1 the non-specific association rate constant [85]. The interesting part is the second term, which gives an “effective transfer rate” [85]. It describes the translocation along DNA towards the target after the first landing on DNA. D_T refers to the total concentration of (non-specific) base pairs and K_{RD} to the non-specific binding constant [85].

Usually the reaction described by the first rate is very fast due to the large number of possible non-specific binding positions, such that it can be considered as an upper limit for k_a [85]. This makes sense, since the target cannot be found faster than the DNA as a whole. The exact form of the second term depends on which translocation mechanisms are taken into account. In the appendix of their seminal paper, Berg and co-workers study a discrete model of the search process [85]. They state the following form of the rate k_2 appearing in the second term on the right hand side of Eq. (2.22) which involves the rate of inter-segment transfer to any site on DNA (ν) and the concentration of operators⁹ (O_T):

$$k_2 O_T = \frac{\Lambda + \nu}{M \tanh(\Omega) \coth(M\Omega) - 1} \text{ with } \Omega = \ln \left(\sqrt{\frac{\Lambda + \nu}{4\Gamma}} + \sqrt{1 + \frac{\Lambda + \nu}{4\Gamma}} \right). \quad (2.23)$$

Γ refers to the sliding rate (the discrete analogue of D_1) and in deriving Eq. (2.23) it was assumed that the fraction Λ/k_{off} of dissociating particles re-associates to a random position along the DNA, while the remaining fraction returns to the dissociation position. The fluxes due to IT and macroscopic dissociation are simply additive, since for ITs it is assumed that they lead to a random position along the DNA.

⁹Note that it is assumed that there is a single binding site on a single DNA molecule within the cell, while in a real bacterial cell there is on average more than one DNA molecule and there are multiple binding sites.

In the specific case when the TF is not able to slide, $\Gamma = 0$, one obtains

$$k_2 O_T = \frac{\Lambda + \nu}{M - 1}. \quad (2.24)$$

If additionally, the TF cannot perform ITs, $\nu = 0$, this corresponds to a search without any facilitating mechanism [85]. The TF lands on a random base pair, remains there and dissociates. In this manner it has to probe (nearly) all non-specific binding positions until it eventually finds the target. It is noteworthy that this finite limit of Eq. (2.23) exists, since in models which *a priori* assume that the target is detected via sliding, as for example Eq. (2.20), the target detection time diverges for a TF which cannot move along DNA (compare also the discussion in [88]).

In the biologically more relevant regime, where the sliding rate is much larger than the dissociation rate or the rate at which an IT loop is formed, $\Gamma \gg \Lambda + \nu$, we have [85]:

$$k_2 O_T = \frac{\Lambda + \nu}{M \sqrt{\frac{\Lambda + \nu}{4\Gamma_1}} \coth \left(M \sqrt{\frac{\Lambda + \nu}{4\Gamma_1}} \right) - 1}, \quad (2.25)$$

in clear analogy to the previous result, Eq. (2.20). It is important to note that this result does not imply *per se* that the target will be found quicker than with three-dimensional diffusion alone. We will discuss this issue below.

2.3.1 How is sliding made possible biologically?

Even if one is able to show theoretically that motion in reduced dimensionality can be beneficial, it is not yet clear if such a motion is possible biologically. How this works out was addressed by Robert B. Winter and co-workers in 1981 [59]. Based on earlier observations that the non-specific interaction is mainly electrostatic and mediated by the release of counter ions, they supposed that sliding happens on an isopotential cylinder surface [59]. This can be understood as follows: during sliding motion counter ions in front of the repressor are displaced, but quickly replaced by others behind the repressor [59]. Under these circumstances, it is expected that the activation barrier which has to be expended for this kind of motion is rather small [59]. Related energy landscapes were later discussed by Dahirel and co-workers [89] and by Bénichou and co-workers [90]. The importance of non-specific binding *in vivo* was highlighted for example by Bakk and Metzler [64].

2.4 Modern studies

After the introduction of the facilitated diffusion model, there came a rather silent phase, in which not many studies on this subject were published. This changed in recent years, mostly due to an enormous progress in experimental possibilities: while in the 1970s and early 1980s one-dimensional sliding of proteins along DNA had been a reasonable assumption, by now it was possible to show that it is an experimental reality [91]. Not only in this respect it has become possible to study biological processes on a single molecule level, see e.g. the comprehensive review by Gene-Wei Li and X. Sunney Xie [25]. We shortly summarise the important recent experimental findings.

2.4.1 Experimental studies

In the 1990s several groups repeated the measurements of the association of lac repressor to the operator, yielding more [92] or less [71] the earlier results. Thereby, each of the experimental groups criticised the techniques that previous studies used. The motion of proteins in the cytoplasm of cells was studied, showing subdiffusion for mRNA [93] and normal diffusion for other proteins [94]. In another context it was shown that mRNA in bacteria can remain close to their site of transcription for rather long times [95]. Other works focused on specific microscopic processes necessary for the facilitated diffusion mechanism (hopping/jumping and importantly sliding): the latter has been unambiguously shown to take place [96–103].

Most importantly, Johan Elf and co-workers performed single-molecule experiments with the lac repressor in living *E. coli* cells. In the first study, the slightly modified lac system was first induced by adding IPTG and subsequently anti-induced by adding 2-nitrophenyl- β -D-fucoside (ONPF) [39]. Then it was monitored at which rate the TFs re-bind to one of the operators in the system. For the three repressors which were typically present in the cell in their experiments, they obtained that the first one of these will on average bind to one of the two operators after 59 s [39]. A single repressor should then find a single operator after roughly $59 \text{ s} \times 2 \times 3 = 354 \text{ s}$ [39]. They also found the residence time on DNA to be in the range of 1 ms [39]. Besides, using an additional *in vitro* assay they could determine a one-dimensional diffusion coefficient of $D_1 = 0.046 (\mu\text{m})^2/\text{s}$ for the repressor.

In a follow-up study, Hammar et al. used different strains of *E. coli* in which the region around the main operator was modified stepwise [104]. Introducing a second strong operator near the target and binding sites for other proteins as roadblocks near the target, they were able to show three main points: first, that the lac repressor uses the facilitated diffusion mechanism in living cells with sliding lengths of $(45 \pm 10) \text{ bp}$ [104]. Second, that roadblocks in the form of other bound proteins cannot be by-passed by the sliding lac repressor [104] (compare the simulations by Marcovitz and Levy [105]). And third, that in about 90% of the cases the repressor will overshoot the target without detecting it [104]. In a very recent paper, Hammar and co-workers compared the binding of lac repressor to the natural O_1 -operator with binding to the symmetric operator O_{sym} , finding that dissociation from the stronger operator is slower, while the association rate is similar in both cases [106].

Going one step further, the group of X. Sunney Xie has published important studies on gene expression: using fluorescent markers, they observed the expression of single proteins [24], showed that a stochastic single-molecule event can determine the phenotype of a cell [18] and finally they quantified the *E. coli* proteome and transcriptome [107]. Mostly based on thermodynamic models and using deep sequencing more and more microscopic details of the lac system can be assessed quantitatively [40, 108, 109]. This equips simulations with reasonable parameter values.

Some experimental as well as theoretical studies focused on the looping of DNA mediated by tetrameric lac repressors which are bound to two operators [110, 111], while others drew conclusions from the organization of genes on chromosomes [112–121]. These works often studied the co-localization effect, i.e. that some DNABPs are produced close to the place where they have to perform their task later on [120]. In turn, the publication

of these experimental studies led to a renewed interest of theoreticians in this subject over the last decade.

2.4.2 Theoretical studies

In recent years an immense number of theoretical studies on the search of DNABPs for a target sequence on DNA was published. This can already be seen by noting that there is a whole series of related review articles, each emphasizing different aspects [72, 122–125]. Therefore we only shortly mention most works, and focus on those which were particularly influential and/or relevant for our studies. We start with those which are based on a stochastic approach. What is common to all of them is that the rebinding position on DNA is assumed to be uncorrelated with the position where the TF dissociated. All two-state models and a motivation of what is meant by that will be presented in subsection 2.4.4.

In 2004 Michael Slutsky and Leonid A. Mirny formulated a simple version of the facilitated diffusion model [126]. Using the notation, $\bar{\tau}_{1d}$ and $\bar{\tau}_{3d}$ for the mean times spent on and off the DNA, they wrote the mean search time as [126]:

$$t_s(\bar{n}, M) = \frac{M}{\bar{n}} [\bar{\tau}_{1d}(\bar{n}) + \bar{\tau}_{3d}], \quad (2.26)$$

where M denotes the number of base pairs in the DNA and \bar{n} the average number of base pairs that is scanned during a sliding event. From this form which is equivalent to Eq. (2.17) (apart from the term τ_1), they deduced the often cited result that the search is optimised when equal amounts of time are spent diffusing in 1D and in 3D: $\tau_{1d}(\bar{n}_{opt}) = \tau_{3d}$ [126]. Quite optimistically they predicted an in vivo search time of a single repressor to be in the range of (1 – 10) s. More realistic was the estimation of a sliding distance of 50 bp based on the assumption of a one-dimensional diffusion coefficient which is two orders of magnitude smaller than the one in 3D [126]. Interestingly, they also noted that the presence of nucleosomes which block the searching particles can even be advantageous [126]. We return to their model in subsection 2.4.4 when we discuss two-state models.

By the time that Mathieu Coppey and co-workers published their paper, Slutsky and Mirny's work was already available as a preprint. They found that when both the dissociation and the re-association times are exponentially distributed (with rates k_{off} and λ'), the Laplace transform of the first passage time density to the target, $\tilde{F}(u)$, can be written as [127]:

$$\tilde{F}(u) = \frac{\langle \tilde{j}(k_{off} + u|x) \rangle_x}{1 - \frac{1 - \langle \tilde{j}(k_{off} + u|x) \rangle_x}{(1+u/k_{off})(1+u/\lambda')}}}, \quad (2.27)$$

where $\tilde{j}(k_{off}|x)$ contains all the information about the one-dimensional search and $\langle \dots \rangle_x$ represents an average of some quantity over the starting position, x , on the DNA. The function $\tilde{F}(u)$ is equivalent to the quantity $\tilde{\Phi}(u)$ appearing in Eq. (2.18). Again, one has $\tilde{F}(u=0) = 1$ implying that all particles will eventually reach the target irrespective of the exact form of $\tilde{j}(k_{off}|x)$. For a sliding particle, and at $u = 0$, this function is given

by [127]:

$$\langle \tilde{j}(k_{\text{off}}|x) \rangle_x = \frac{1}{M+L} \left(2r + \ell \left[\tanh\left(\frac{L-r}{\ell}\right) + \tanh\left(\frac{M-r}{\ell}\right) \right] \right), \quad (2.28)$$

where again $\ell = \sqrt{D_1/k_{\text{off}}}$. Here $2r$ denotes the size of the target, and $L-r$ and $M-r$ the flanking lengths of non-specific DNA to the left and to the right of the target, such that $L+M$ is the total length of DNA. The finiteness of the target enables a direct target detection from the bulk solution as implied by the first term in brackets on the right hand side of Eq. (2.28). This is a feature which many theoretical studies lack, as criticised by Veksler and Kolomeisky [88]. The second term in brackets denotes the flux of particles which reach the target via sliding. For the mean time of target association, one obtains [127]:

$$\tau = \left(\frac{1}{k_{\text{off}}} + \frac{1}{\lambda'} \right) \left(\frac{1}{\langle \tilde{j}(k_{\text{off}}|x) \rangle_x} - 1 \right). \quad (2.29)$$

In the case of an infinitely thin target, $r = 0$, which can only be found via sliding and which lies in the middle of the DNA chain, $M = L$, one has

$$\langle \tilde{j}(k_{\text{off}}|x) \rangle_x = \frac{\ell}{L} \tanh\left(\frac{L}{\ell}\right). \quad (2.30)$$

Plugging this form into Eq. (2.29), one obtains a mean search time which is equivalent to Eqs. (2.17) and (2.26), where the first bracket contains the mean times spent in 1D and 3D and the second bracket the number of search rounds.

Importantly, for long DNA chains and an infinitely small target, Coppey and co-workers were able to show that the dissociation rate which minimises the search time is given by [127]:

$$k_{\text{off},\text{min}} = \lambda' - \frac{4\sqrt{D_1\lambda'}}{L+M} + \mathcal{O}((L+M)^2). \quad (2.31)$$

This generalises Slutsky and Mirny's simpler result that an equipartition of times is optimal [126]. As can be seen from Eq. (2.31) this is only valid for very long chains. Considering a very long chain, but a finite target size, this results becomes [127]:

$$k_{\text{off},\text{min}} = \frac{(\lambda'r + \sqrt{\lambda'r^2 + D_1\lambda'})^2}{D_1}. \quad (2.32)$$

In this equation it becomes apparent that for rather large target sizes and/or slow one-dimensional diffusion the optimal dissociation is larger than the re-association rate. Thus, the less effective one-dimensional search is, the sooner this search mode should be left. This underlines the statement that the "facilitated" diffusion model does *not* facilitate the search for all microscopic parameters.

Subsequently, this approach was generalised by Eliazar and co-workers in order to include anomalous transport mechanisms [128, 129]. Another route of generalisation was realised by Bénichou and co-workers when the assumption of complete loss of correlation between dissociation and re-association was abandoned. The calculated jump distributions were compared with experiments performed with the restriction enzyme

EcoRV [99,130]. Their two-state model introduced in [90] and extended in [124] will be discussed below.

Alexander Y. Grosberg and co-workers wrote a series of papers shedding light on different aspects of the target search process. These authors studied the question what happens when the non-specific binding energies are not homogeneous [131]. In another work the role of inter-segment transfer was discussed [132]. IT was also studied as a means of translocation along DNA by other authors, see e.g. [133], where the problem was treated in terms of Lévy flights. Finally, Grosberg and co-workers focused on the role of the conformation of DNA [134], and showed an intriguing analogy of search processes to electrostatic problems. Their result invoked a large number of different scaling regimes. Interestingly, just like Slutsky and Mirny they found that the optimal strategy implies to spend equal amounts of time on and off the DNA [134].

The dependence of the target association rate on the DNA's conformation was also studied by other authors, namely the group of Andrew J. Spakowitz [135,136] and by our group where the experimental results of van den Broek et al. [137] were described by the model presented in Lomholt et al. [138]. In the experiment of van den Broek et al. a dual optical tweezers set-up was used to study the association of the restriction enzyme EcoRV with a single DNA molecule [137]. While van den Broek et al. found that target association proceeds faster in a random coil than for a target on straight DNA, the gist of the theoretical and computational study of Koslover et al. was that the geometry does not play a decisive role [136,137]. More details will be considered in chapter 4 which is devoted to this issue. The role of dynamically varying DNA conformation was studied in [136,139]. Since in a living cell, there are way more other macromolecules than in an *in vitro* experiment, these crowding effects came into attention. This was detailed in theoretical studies [140–142] as well as in simulations (see below).

Extending the FPT approach of Szabo and co-workers, Konstantin V. Klenin and co-workers recalculated the mean target search time under the assumptions that at all distances from the target the one-dimensional and the three-dimensional pathways are equilibrated [143]. They obtained a result which looks very similar to the one for a target with the form of an oblate spheroid obtained by Richter and Eigen [82]. This was also compared to a computational model presented in [144].

Similarly, Cherstvy et al. introduced a “correlation” term between the one-dimensional and the three-dimensional portion of a single search round [145]. Importantly, in the second half of this paper also the recognition of the target sequence was modelled. In the spirit of Berg and Ehrenberg, Huan-Xiang Zhou studied the search process for a target of finite extension on a straight cylinder in terms of a steady-state Smoluchowski equation involving a surface potential [146–148]. Further theoretical models were introduced by Belotserkovskii and Zarling [149], Park et al. [150] and Malherbe and Holcman [151].

2.4.3 Computational models

In parallel to the theoretical studies, the enormous increase in computational power led to the publication of many computational models. Nicolae-Radu Zabet wrote a whole series of papers dealing with gene regulation in prokaryotes [152–158].

Ana-Maria Florescu and Marc Joyeux showed in three papers that the facilitated diffusion mechanism took place in their simulations, but argued that its potential to

yield faster-than diffusion rates in living cells is limited [73, 159, 160]. We will comment on these criticisms in subsection 2.5.2.

Davide Marenduzzo and co-workers published two Monte Carlo studies on this subject. The first one focused on the influence of different types of confinement and of several values of the stiffness of DNA on the search dynamics [161]. Most importantly, they also studied different concentrations of searches, an aspect which is often overlooked, and found that in the case of ten searchers variations of the non-specific affinity around the optimal value do not change the search time drastically [161]. Their second publication introduced other proteins as “crowders” and “blockers” into the system [162]. Again they found that in their simulations which mimic the crowded interior of the cell the search time is rather “robust” [162]. This nicely agrees with earlier experimental findings [163] and we will find out in chapter 4 that our semi-analytical model is able to recover this robustness.

The group of Yaakov Levy performed several numerical studies on protein-DNA search which focused on different aspects on the problem. Lately, this was the influence of mobile and immobile blockers for the search process on a 100 bp long stretch of DNA [105]. Finally, Yann von Hansen et al. studied the influence of hydrodynamic effects on DNA-protein binding [164].

2.4.4 Two-state models and the speed-stability paradox

Already in 1981 it was hypothesised by Robert B. Winter and co-workers that there might be something like a specific and a non-specific binding mode. They described the sliding of TFs as “movement along a continuous (locally ‘featureless’) cylinder” [59]. Furthermore, they supposed that unlike in binding to the target where specific hydrogen bonding takes place, faced with non-specific DNA the TF performs a change of conformation, in which the hydrogen bonding groups are no longer present, but in which the electrostatic interaction is maximised [59]. Importantly, they formulated the main requirement for an efficient search process [59]:

“a general [...] affinity between the ‘target-seeking’ moiety and the general (nonspecific) target surface and a rapid equilibrium between the target-specific form and a target-nonspecific form of the binding ligand.”

However, these two states were not considered explicitly in their theoretical model. Some modern models do this because of the so-called *speed-stability paradox*, sometimes also referred to as speed-selectivity paradox. What is meant by that?

Based on the quite reasonable assumption that the protein-DNA binding energy can be well approximated by a Gaussian distribution, Slutsky and Mirny interpreted the sliding motion as a walk in a random energy landscape. As shown in the late 1980s by Robert Zwanzig, if the roughness of the potential is denoted by σ , the ensuing diffusion coefficient is given by $D \propto \exp(-\sigma^2)$ [165]. In terms of the search process this means that fast sliding requires the roughness of the binding energy profile to be at most in the range of $1 k_B T$ [126]. Conversely, since the target site is by definition the site with the largest binding energy, it can be shown that a sufficiently high stability implies that the roughness should be approximately $5 k_B T$ for a genome with a size in the order



Figure 2.4: Scheme of the two conformational states of a TF. While in the search state (red), the particle is able to slide quickly along DNA, but it cannot detect the target. In the recognition state (blue), the particle’s mobility is reduced, but it is able to bind specifically to the target.

of magnitude of 1 Mbp^{10} [126]. Obviously, if there is just one binding mode these two constraints cannot be fulfilled at the same time.

Therefore the existence of two different binding modes was supposed. In the first one, often termed “search” state, the protein is only weakly bound and experiences a rather smooth energy landscape which enables fast sliding [126]. This is illustrated by the red particle in Fig. 2.4. The second state, named “recognition” mode, conveys a much more rugged energy landscape, but it is in this mode that the particle binds specifically [126]. Compare the blue particle in Fig. 2.4.

Slutsky and Mirny also supposed that the conformational change between these two states involves partial folding of the TF and extrusion of water [126]. For LacI, such considerations were put on stable experimental grounds by NMR studies which showed that a helix is formed in the hinge region in the specifically bound repressor, but not in the non-specifically bound one [166]. Thus, at least this change in the hinge region must take place when a conformational change occurs. Kalodimos et al. further stated that such motions happen on the micro- or milli-second time scale [166].

Importantly, Slutsky and Mirny proposed that the rate at which conformational changes occur is coupled to the underlying nucleotide sequence, in such a way that they are more probable to occur at sites which are similar to the target site [126]. The decisive role of the barrier height associated with the switching between conformations was highlighted: if it is too low, at too many sites the DNA will be checked for the target, thus rendering the two-state scheme ineffective. Conversely, if it is too high it becomes increasingly probable to miss the target [126].

Before we continue with a descriptions of other two-state models, it is time for a cautionary remark. It was noticed early on by Robert B. Winter and co-workers, that mapping the conformation of the protein on the two extreme states is just an approxi-

¹⁰Later these calculations were extended by Sheinman et al. who took also gapped energetic distributions into account [124].

mation to the real continuum of possible binding interactions [59]. Thus, such two-state models should only be considered to be a working hypothesis.

The contribution of Longhua Hu and co-workers to the two-state models was particularly important because they noted that effects similar to “impedance matching” are at work [167]. While in the case of an infinitely fast target detection, the one-dimensional search phase can be optimised by increasing the one-dimensional diffusion coefficient, for a finite target detection rate an increasing of the speed of one-dimensional diffusion brings along the risk of “overshooting” the target. This is important in light of the already mentioned experimental findings which implied that the target will not be recognised at every encounter [104]. Thus, simpler one-state models might not be able to describe the situation appropriately. Finally, Hu et al. assumed that the interconversion rates between the two states is fast enough to consider the system as equilibrated.

We will not detail the model described by Reingruber and Holcman, since it will be contained as a limiting case of our model presented in chapter 3 [168, 169]. However, we report here one of their central findings which at first sight is counter-intuitive: they considered an one-dimensional interval in which a particle diffuses and meanwhile switches randomly between two states. One of the states is characterised by a higher diffusion coefficient, but in this state the target at the left end of the system cannot be found. Conversely, in the slower diffusing state the target at the left end can be detected. What they found was that this happens the fastest if most of the time is spent in the faster state in which, however, the particle is target-insensitive [168]. This is in clear contrast to the equilibrium assumption of Hu et al. [167].

Olivier Bénichou and co-workers introduced a two-state model which was later extended [90, 124]. Building on the stochastic approach of Coppey et al., they assumed that probing for the target occurs rarely in comparison to sliding or dissociation. They obtained that there are two characteristic time scales, a short one in which the recognition state is never entered at a non-target site and a long one in which this happens. This leads to differences between typical search times and the mean search time, an effect which becomes increasingly important when several proteins are searching for the target [90].

Further two-state models were described by Rajamanickam Murugan [170] (and references therein) and Huan-Xiang Zhou [171] who both extended their previous one-state models and lately by Shi Yu and co-workers [172].

2.5 General features of the model and criticism

In this final section of the present chapter, we conclude what is common to all facilitated diffusion models. Afterwards we consider the criticism which has been uttered towards it. As a general caveat we note that the values of most of the microscopic parameters which enter the model are not known exactly. Since these quantities are multiplied with each other, final results should be taken with a grain of salt, compare [67].

2.5.1 General features

The facilitated diffusion model is based on the observation that many DNA-binding proteins have a general affinity for DNA. That this is possible at all is due to its ability

to specifically bind its target [125]. Thus, although some proteins may bind solely to non-specific DNA stretches without having a specific target, it is simply impossible to have a protein which binds specifically to a target without having a non-specific binding affinity [125].

In terms of the search process alone, this constitutes a liability because each non-specific binding position effectively acts as a trap. It only pays off if the protein is able to actively use this “junk DNA”, either by sliding along it or by invoking inter-segmental transfer. In terms of sliding, the benefits can be attributed to the “antenna effect” mentioned before: the antenna consists of the range of non-specific binding positions near the target from which the latter can be detected via sliding. Thus, for the searching particle the target appears to be larger. However, the facilitated diffusion model only accelerates the search if this antenna effect is able to outperform the slow-down due to binding to other non-specific sites far away from the target. These effects clearly depend on the concentration of non-specific sites: when the concentration of non-specific DNA is very high the search process can become very ineffective.

Another subtle effect is the “tethering of proteins to the DNA” [136]: in dilute conditions, being bound to non-specific DNA implies that at least the DNA as a whole has been found and the effective protein concentration near the target increases. This can be considered to be the little sibling of the increase of local concentration due to looping.

A further important point is the concentration of searching particles. If their concentration is high, the time interval between events when one of the searchers lands near the target will be short. Then the possible acceleration due to long sliding lengths will become increasingly unimportant. This is particularly true for RNAP, which has copy numbers of a few thousand in a living cell, while acceleration effects are more important for sparsely present TFs like the lac repressor [125]. Concentration effects were for example studied by Sokolov and co-workers showing that for purely three-dimensional search the target detection rate is directly proportional to the concentration of searching particles, while for a purely one-dimensional process this rate is proportional to the square of the concentration [97].

2.5.2 Modern criticism

As a popular theoretical approach, the facilitated diffusion model also faces some criticism which should be mentioned here. Some of the authors criticise the often heard statement, that the facilitated diffusion mechanism enables reactions above the diffusion limit. Most notably, Stephen E. Halford wrote a critical article entitled “An end to 40 years of mistakes in DNA-protein association kinetics”, even though he himself derived a simplified version of the facilitated diffusion model several years before [72, 173]. The central point of his critique is that the measured high association rates are not truly faster than the diffusion limit when taking into account the electrostatic interaction. Consequently, in his way of speaking it is the diffusion limit itself which varies with the salt concentration [72]. However, since there are so many non-specific sites to which the TF can bind, he makes a statement which is close in spirit to Riggs et al.’s interpretation of the association rate [72]:

“[...] the surprising feature of DNA-protein association is not that they are

often rapid but rather that they are not incredibly slow.”

This is the central point. Often the observed target association rate is directly compared with the corresponding rate in the hypothetical situation without interaction with non-specific sites. Obviously, in the crowded interior of a living cell it could be faster to “switch off” the interaction with non-specific sites, just to diffuse in 3D and to associate with the target. However, a real protein does not have this option. As detailed above, the ability to form a specific complex with the target site implies that there is a non-specific interaction, at least to some extent [125]. Thus, it seems more appropriate to compare association rates of a full model with the ones obtained with a minimal model in which non-specific interaction occurs, but no sliding is possible. Then, however, any sliding ability will speed up the process. In general, in a cell non-specific binding keeps the TFs close to the genome and increases their local concentration.

Furthermore, it has been argued that facilitated diffusion might not be necessary for all proteins, especially when they come in high copy numbers. While this is certainly true, the statement that a fast search process could be enforced by increasing the copy numbers of the searching proteins remains questionable [125]. In a living cell many proteins have to find their target sequences. If all of them “chose” to increase their copy numbers, this would aggravate crowding effects; most importantly, their self-diffusion would slow down and it would be questionable if such a search process still was more efficient. Thus, any study invoking more searching particles should take such effects into account. For example, Gene-Wei Li and co-workers made the simplifying assumption that the searching proteins constitute a fixed fraction of the total number of proteins [141]. Increasing their concentration helps if they are present in low copy numbers, but above a certain threshold crowding effects slow down the search such that there is an optimal crowding level in between these two regimes [141]. Furthermore, one has to note that it is costly to produce TFs.

Alex Veksler and Anatoly B. Kolomeisky questioned the very existence of the speed-stability paradox [88]. In their opinion, models in which the coordinate along the DNA axis is treated as continuous are only approximate and should be replaced by discrete models. Then, considering scanning lengths below 1 bp would yield near-optimal search times. However, this argumentation remains elusive: on the one hand, clearly discrete models are an approximation to the continuous reality and furthermore numerous experiments show that sliding occurs on much longer length scales. Another one of their points of critique is that continuum models with an infinitely small target disregard the possibility of finding the target via 3D diffusion and thus assume a priori target location by sliding. This is definitely true, but as shown already by Coppey et al. continuum models can be combined with targets of finite extension [127].

All in all, in the theoretical description of the association of proteins with target sequences on DNA the facilitated diffusion model is well established.

3 Generalised facilitated diffusion model

The question how proteins are able to bind quickly to target sequences on DNA is of paramount importance. We deal with this problem by studying the association of a single protein with a single target. The model presented in this chapter is generic, but whenever we perform explicit numerical calculations we will use values obtained for the lac repressor in *E. coli*, because this is the system which is best characterised quantitatively. The approach described herein is based on the formulation of the search problem that was presented in Otto G. Berg's first paper on this topic in 1976 [84]. However, our work goes beyond this early approach by combining this search model with the assumption that the protein can be present in two different conformations while it is bound to DNA. This two-state model was anticipated by Winter and co-workers in 1981 and formally introduced by Slutsky and Mirny as already described in detail in subsection 2.4.4 [75, 126].

The combination of both approaches is very helpful, because it allows a closed-form solution of the problem. Thereby, the central quantity to be calculated is the mean search time, τ . Its knowledge enables us to obtain what is usually measured experimentally, the specific association rate, k_a which is given in units of $\text{M}^{-1}\text{s}^{-1}$:

$$k_a = \frac{1}{\tau n_0}. \quad (3.1)$$

Again the operator density on the DNA is denoted by n_0 . We now first describe the model in detail, before we present the results in sections 3.2 and 3.3. Section 3.4 relates this model to previous approaches and section 3.5 will conclude this chapter by describing its benefits and shortcomings.

The content of this chapter is an comprehensive extension of the work published before in [174].

3.1 The model

In this model, which we call *generalised facilitated diffusion model* (GFDM), the DNA is considered as a cylinder characterised by its radius R_1 and its contour length $2L$. Both the situation of a straight cylinder and of a completely coiled configuration will be considered. For straight DNA we use the closed-cell approach which was already mentioned in section 2.2.3. It was noticed early that the difference to the related approach in which the TF completely loses its correlation with the unbinding position at a certain radial distance is modest [86]. In all cases, the operator which will be called target in the following, is assumed to be infinitely thin and in the middle of the DNA, such that the flanking length on both sides of the target amounts to L .

The situation is illustrated in Fig. 3.1, where for the sake of visibility a target of finite extension is shown in form of the black part of the cylinder's surface. The TF is



Figure 3.1: Scheme of the search for a target (black stretch) on a straight cylinder in terms of the facilitated diffusion model. The TF lands on DNA, slides along it and dissociates again.

represented by the red U-shape. The partially transparent shapes show previous positions of the TF.

As initial condition we put the searching TF at a random position in the bulk solution. Note that this is not an equilibrium starting configuration, but is motivated by a typical experimental set-up in which the timer starts after fully inducing the system, when all TFs are unbound [39]. Of course many other experimental set-ups are possible, see for example reference [137].

The TF diffuses with diffusivity D_3 until it encounters a stretch of DNA. Upon such an encounter it can bind to the DNA with the non-specific association rate k_{on}^1 . This is shown in the left half of Fig. 3.1. While bound to the DNA, it can slide along the latter which is usually modelled as one-dimensional diffusion. As a side remark we note that there are also some studies which model it as a two-dimensional random walk on the DNA surface [176]. Importantly, this is not directed motion as implied by the overlapping previous positions of the TF on DNA in Fig. 3.1.

When bound to the DNA the two different states of the TF play a role. In the loosely bound state (“search state”), the one-dimensional diffusion coefficient D_s will be much larger than the one in the tightly bound “recognition state”, D_r . The reason for that is that we assume that the binding energy profile along the chain is much more rugged in the recognition state. Formally this amounts to

$$D_i \propto \exp(-\sigma_i^2), \quad (3.2)$$

with σ_i , the variance of the binding potential between DNA and the TF, where $i = s$ refers to the search state and $i = r$ to the recognition state [126, 165].

While the TF is non-specifically bound, apart from sliding to neighbouring positions there are two further options:

- dissociating from the DNA with rate k_{off} or

¹Note that there are different conventions in which dimensions such a binding rate is defined [175]. In the present case k_{on} has units of length per time unit.

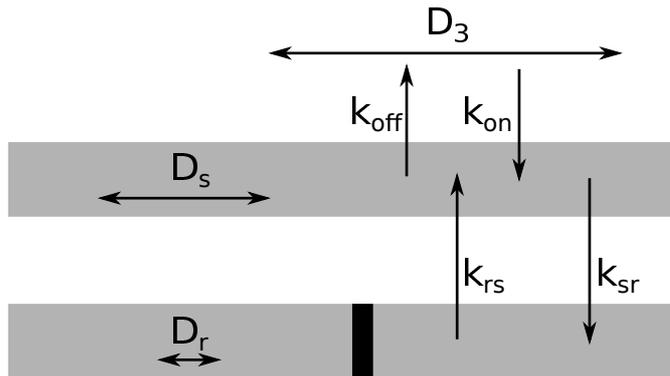


Figure 3.2: Scheme of the processes which make up the GFDM. In the bulk solution the TF moves with the three-dimensional diffusion coefficient D_3 and binding to DNA in the search state (upper rectangle) occurs on encounter with the microscopic rate k_{on} . In the search state the TF diffuses with the one-dimensional coefficient D_s and dissociates with rate k_{off} . While in the search state, with rate k_{sr} the TF changes its conformation to the recognition state. Only in this state, in which the TF diffuses with a one-dimensional coefficient D_r , the target (dark stripe) can be found. With rate k_{rs} the TF returns to the search state. Note that search and recognition state simply reflect different modes of the TF and positions within the two rectangles shown in this scheme should not be considered as different positions in real space.

- making a configurational transition to the recognition state with rate k_{sr} .

Note that for the sake of simplicity Fig. 3.1 only shows a dissociation event, but no switching to recognition mode which was shown in blue in Fig. 2.4. The distinction between the two conformations is more clearly illustrated in the schematic figure 3.2, in which all dynamic processes within the GFDM are shown.

When the TF is bound in the recognition mode, we assume that the binding is so strong that no direct dissociation is possible. Consequently, besides sliding with the (reduced) diffusivity D_r , the only option for the TF in the recognition mode is to return to the search state with rate k_{rs} . However, unlike in the search state a TF in the recognition mode is able to detect the target. Within this model we assume that in the search mode the TF is completely unaware of the underlying sequence such that switching to the recognition mode happens randomly. In particular, there is no *induced switch* to the recognition mode at the target site. Such effects will be discussed in chapter 5.

Now we specify the quantities we examine in the following: the one-dimensional probability densities (per length) of the TF on the DNA, $n_i(z, t)$, where again $i = r, s$. Based on the above discussion they obey the following fundamental set of coupled differential equations, which are supplemented with boundary values below. First, for particles in the recognition mode:

$$\frac{\partial n_r(z, t)}{\partial t} = D_r \frac{\partial^2 n_r(z, t)}{\partial z^2} - k_{\text{rs}} n_r(z, t) + k_{\text{sr}} n_s(z, t). \quad (3.3)$$

Second, for particles in the search mode:

$$\begin{aligned} \frac{\partial n_s(z, t)}{\partial t} = & D_s \frac{\partial^2 n_s(z, t)}{\partial z^2} + k_{rs} n_r(z, t) - k_{sr} n_s(z, t) + G(z, t) - k_{\text{off}} n_s(z, t) \\ & + k_{\text{off}} \int_0^L dz' \int_0^t d\tau F(z, z', t - \tau) n_s(z', \tau). \end{aligned} \quad (3.4)$$

In the upper Eq. (3.3), we recognise that particles in the recognition mode can slide slowly or convert to become a particle in the search mode. Additionally, according to the last term whenever there is a particle in the search mode at the corresponding position, it can also convert to become a recognising particle.

In the lower Eq. (3.4), we note terms that are equivalent to the ones just mentioned. Apart from that, particles in the search state are allowed to dissociate ($k_{\text{off}} n_s(z, t)$) and there are two terms involving kernel functions. One of them, $G(z, t)$, denotes the so-called ‘‘virgin flux’’, a term introduced in [138]. This is the influx of particles which have not yet been bound to the DNA. The other one, appearing in the last term of Eq. (3.4) describes particles which have dissociated from the chain at an earlier time and now return to the DNA. Note that we do not include IT in our model, but this could be included for the sake of completeness in future studies.

Before we continue the discussion by specifying the initial conditions, we have to admit that the approach is not without problems: in this model the sliding motion appears in a continuous one-dimensional space. Thus, the target has no finite extension implying that it is not possible to detect the target directly from the bulk solution. *A priori*, we assume that target detection occurs by sliding and we cannot expect to obtain meaningful results in the limit when D_r approaches zero. Thus, within the model no direct comparison to a purely three-dimensional search can be made.

Since we assume that initially the particle is unbound, at $t = 0$ there is no probability density in either state on the DNA: $n_r(z, t = 0) = n_s(z, t = 0) = 0$. This suggests to transform the problem to Laplace space with respect to time, yielding the following two equations:

$$\left(u + k_{rs} - D_r \frac{\partial^2}{\partial z^2} \right) \tilde{n}_r(z, u) = k_{sr} \tilde{n}_s(z, u), \quad (3.5)$$

$$\left(u + k_{sr} + k_{\text{off}} - D_s \frac{\partial^2}{\partial z^2} \right) \tilde{n}_s(z, u) = k_{rs} \tilde{n}_r(z, u) + k_{\text{off}} \int_0^L dz' \tilde{F}(z, z', u) \tilde{n}_s(z', u) + \tilde{G}(z, u). \quad (3.6)$$

To solve these equations, we need to supply some boundary conditions. First, we note that due to symmetry reasons, it suffices to consider positive values of z , i.e. only what happens in the right half of the DNA. Then, for the particle in the search mode both $z = 0$ and $z = L$ act as reflecting boundaries, as was shown experimentally for the endonuclease *EcoRV* [177]:

$$\left. \frac{\partial n_s}{\partial z} \right|_{z=0} = \left. \frac{\partial n_s}{\partial z} \right|_{z=L} = 0. \quad (3.7)$$

For particles in the recognition mode, the boundary at $z = L$ is also perceived as reflecting, however, at $z = 0$ the particles will be absorbed by the target:

$$\left. \frac{\partial n_r}{\partial z} \right|_{z=L} = 0 = n_r(z = 0, t). \quad (3.8)$$

The second part of Eq. (3.8) implies that in the recognition state the particle detects the target with a probability of 100% when it slides to its location.

Since the mean first-passage time to the target, τ , is given by

$$\tau = - \left. \frac{\partial \tilde{j}(u)}{\partial u} \right|_{u=0}, \quad (3.9)$$

we first have to determine the Laplace transform of the flux of particles to the target, $\tilde{j}(u)$, which appears on the right hand side of Eq. (3.9). Since by assumption solely particles in the recognition mode have the ability to detect the target, in the time domain this flux is related to their density via

$$j(t) = D_r \left. \frac{\partial n_r(z, t)}{\partial z} \right|_{z=0}. \quad (3.10)$$

To solve this equation explicitly, two steps have to be made: first the kernel functions F and G which describe the three-dimensional motion in the bulk solution have to be determined and then these results have to be plugged into Eqs. (3.5) and (3.6).

3.1.1 3D kernel functions for the straight DNA conformation

In the following we derive the solution of F and G for straight DNA chains arranged in parallel. The case of completely coiled DNA will be discussed in section 3.3.

In the bulk solution the TF undergoes normal diffusion. Thus, due to the symmetry of the problem (rotational symmetry around the axis of the cylinder, which coincides with the z -axis), we can again focus on positive values of z and one has to solve the following cylindrical diffusion equation:

$$\frac{\partial P(r, z, t)}{\partial t} = D_3 \left(\frac{\partial^2}{\partial z^2} + \frac{1}{r} \frac{\partial}{\partial r} r \frac{\partial}{\partial r} \right) P(r, z, t). \quad (3.11)$$

According to the closed-cell approach, we assume the outer extremities of the system to be reflecting boundaries [84]. Technically, this implies

$$\left. \frac{\partial P(r, z, t)}{\partial z} \right|_{z=0, L} = \left. \frac{\partial P(r, z, t)}{\partial r} \right|_{r=R_2} = 0. \quad (3.12)$$

Determination of F As mentioned above, the function F describes the return of particles to DNA which have dissociated before. In Laplace space, \tilde{F} is given by

$$\tilde{F}(z, z', u) = 2\pi R_1 k_{\text{on}} \tilde{P}(r = R_1, z, u), \quad (3.13)$$

where \tilde{P} denotes the solution of the cylindrical diffusion equation, Eq. (3.11), subject to the initial condition, $P(r, z, t = 0) = 0$, and the additional boundary condition

$$D_3 \left. \frac{\partial P(r, z, t)}{\partial r} \right|_{r=R_1} = -\frac{\delta(t)\delta(z-z')}{2\pi R_1} + k_{\text{on}} P|_{r=R_1}. \quad (3.14)$$

Here the second term involving the two delta functions describes a particle which dissociated at $t = 0$ and augments the rest of the equation which describes just a reactive boundary at the inner cylinder surface. The calculation proceeds analogously to the one in Berg and Blomberg [84]. One obtains the expansion

$$\tilde{F}(z, z', u) = \sum_{m=0}^{\infty} g_m(u) \cos(m\pi z/L) \cos(m\pi z'/L). \quad (3.15)$$

Here the auxiliary function, $g_m(u)$ involves the term, $a_m = u/D_3 + m^2\pi^2/L^2$, and is defined via

$$g_m(u) = \frac{2 - \delta_{m,0}}{L} \frac{k_{\text{on}} \Delta_{01}(\sqrt{a_m} R_1)}{k_{\text{on}} \Delta_{01}(\sqrt{a_m} R_1) - D_3 \sqrt{a_m} \Delta_{11}(\sqrt{a_m} R_1)}. \quad (3.16)$$

The symbols Δ_{01} and Δ_{11} are defined similarly to Chechkin et al., where the general problem of bulk-mediated transport along a cylinder was studied [178]. They denote combinations of modified Bessel functions:

$$\Delta_{01}(\sqrt{a_m} r) = I_0(\sqrt{a_m} r) K_1(\sqrt{a_m} R_2) + I_1(\sqrt{a_m} R_2) K_0(\sqrt{a_m} r), \quad (3.17)$$

$$\Delta_{11}(\sqrt{a_m} r) = I_1(\sqrt{a_m} r) K_1(\sqrt{a_m} R_2) - I_1(\sqrt{a_m} R_2) K_1(\sqrt{a_m} r). \quad (3.18)$$

Determination of G The function G is determined similarly via

$$\tilde{G}(z, u) := 2\pi R_1 k_{\text{on}} \tilde{P}(r = R_1, z, u). \quad (3.19)$$

Here \tilde{P} denotes again the solution of the cylindrical diffusion equation, but now subject to the conditions

$$P(r, z, t = 0) = \frac{1}{\pi L (R_2^2 - R_1^2)}, \quad \text{and} \quad (3.20)$$

$$D_3 \left. \frac{\partial P(r, z, t)}{\partial r} \right|_{r=R_1} = k_{\text{on}} P|_{r=R_1}. \quad (3.21)$$

Unlike in Eq. (3.14), in Eq. (3.21) there is no term involving two delta functions, since in this setting the particles start at a random position in the bulk solution as implied by Eq. (3.20). Performing the calculation, one obtains:

$$\tilde{G}(u) = \frac{2R_1}{L(R_2^2 - R_1^2)} \frac{k_{\text{on}}}{u} \frac{D_3 \sqrt{a_0} \Delta_{11}(\sqrt{a_0} R_1)}{D_3 \sqrt{a_0} \Delta_{11}(\sqrt{a_0} R_1) - k_{\text{on}} \Delta_{01}(\sqrt{a_0} R_1)}. \quad (3.22)$$

3.1.2 Solution of the differential equations

With the explicit forms of F and G at hand, the fundamental set of differential equations, Eqs. (3.5) and (3.6) can be solved. Analogously to Berg and Blomberg, this proceeds with the following ansatz for the Laplace transformed probability densities [84]:

$$\tilde{n}_r(u, z) = \sum_{n=0}^{\infty} f_n(u) \sin\left(\frac{2n+1}{2L}\pi z\right), \quad (3.23)$$

$$\tilde{n}_s(u, z) = \sum_{m=0}^{\infty} h_m(u) \cos\left(\frac{m\pi z}{L}\right). \quad (3.24)$$

In principle, we are only interested in the functions f_n , since the flux to the target is given by

$$\tilde{j}(u) = \frac{D_r \pi}{2L} \sum_{n=0}^{\infty} (2n+1) f_n(u). \quad (3.25)$$

We introduce the following auxiliary functions which are denoted by Greek letters and which all have the physical dimension of a frequency: $\alpha_m(u) = u + k_{\text{off}} + k_{\text{sr}} + D_s m^2 \pi^2 / L^2$, $\beta_n(u) = u + k_{\text{rs}} + D_r (2n+1)^2 \pi^2 / (4L^2)$, and $\gamma_m(u) = u + k_{\text{rs}} + D_r m^2 \pi^2 / L^2$. In the following we omit the explicit dependence of all functions on u . Appropriately modifying Eqs. (3.5) and (3.6), we obtain:

$$\frac{2L}{\pi} \gamma_m Y_m + \tilde{j} = k_{\text{sr}} L_m h_m \quad (3.26)$$

$$h_m L_m (\alpha_m - k_{\text{off}} L_m g_m) = G L \delta_{m,0} + \frac{2L}{\pi} k_{\text{rs}} Y_m \quad (3.27)$$

$$\tilde{j} = \frac{2D_r}{L} \sum_{m,n=0}^{\infty} \frac{k_{\text{sr}} h_m}{\beta_n} \frac{2n+1}{(2n+1)^2 - 4m^2}, \quad (3.28)$$

where we additionally introduced the m -dependent length scale $L_m = L(1 + \delta_{m,0})/2$ and $Y_m(u) = \sum_{n=0}^{\infty} f_n(u) \frac{2n+1}{(2n+1)^2 - 4m^2}$. Finally, eliminating Y_m , one obtains

$$\tilde{j} = L G / \left\{ \frac{k_{\text{rs}}}{\gamma_0} + \frac{\epsilon_0 (\gamma_0 (\alpha_0 - k_{\text{off}} L g_0)) - k_{\text{rs}} k_{\text{sr}}}{\gamma_0} \times \left[\frac{L^2}{2D_r k_{\text{sr}}} + \sum_{m=1}^{\infty} \frac{2k_{\text{rs}}/\epsilon_m}{\gamma_m (\alpha_m - k_{\text{off}} \frac{L}{2} g_m) - k_{\text{rs}} k_{\text{sr}}} \right] \right\}, \quad (3.29)$$

where $\epsilon_m(u) = \left(\sum_{n=0}^{\infty} \frac{1}{\beta_n} \frac{(2n+1)^2}{(2n+1)^2 - 4m^2} \right)^{-1}$.

This result for \tilde{j} looks rather involved. However, as we will see below the resulting mean target search time can be brought into a clear form.

3.2 Results of GFDM for straight DNA

In this section we present the result obtained with the kernel functions which were derived for the case of straight DNA chains arranged in parallel. Before we explain the term structure of the mean search time, we first state our reference set of parameters. They are mostly motivated by experimental results. Following the classical experiment of Riggs and co-workers we study an operator concentration of $n_0 = 10^{-12}$ M [55]. With $L = 8 \mu\text{m}$, this corresponds to $R_2 \approx 5.75 \mu\text{m}$ via the relation $n_0 = 1/(2\pi LR_2^2)$. We fix $D_3 = 50 \frac{(\mu\text{m})^2}{\text{s}}$ and $k_{\text{on}} = 5.56 \times 10^4 \frac{\text{m}}{\text{s}}$ and the DNA-protein interaction radius, R_1 , is assumed to be 6 nm [77, 84].

The rates at which the TF switches between its two conformations are not known experimentally. However, some repressor isomerisations are known to occur at rates of approximately $10^4/\text{s}$ [179], compare also [166]. Thus, we use $k_{\text{sr}} = k_{\text{rs}}/10 = 10^4/\text{s}$, guaranteeing that the TF spends most of the time that it is bound in the fast search state [171]. This assumption has been criticised by Veksler and Kolomeisky, since it seems to be counter-intuitive that the search state should be occupied more often and thus more stable than the recognition state [88]. However, this statement is only supposed to be true at an average non-target site. In fact, the two-state model would be hopelessly ineffective if the recognition state was more favourable at all sites. Related energy landscapes have been discussed by Gerland et al. and by Mirny et al. [62, 122].

For choosing appropriate values for the two one-dimensional diffusion coefficients, it is important to note that an experiment with a finite time resolution will yield neither D_s nor D_r , but an apparent diffusion coefficient, D_a given by

$$D_a = \frac{D_s/k_{\text{sr}} + D_r/k_{\text{rs}}}{1/k_{\text{sr}} + 1/k_{\text{rs}}} \approx \frac{D_s}{1 + k_{\text{sr}}/k_{\text{rs}}}, \quad (3.30)$$

where we used the assumption $D_r \ll D_s$ in the second step. In the following we take $D_s = 0.05 \frac{(\mu\text{m})^2}{\text{s}}$, and $D_r = D_s \exp(-\chi)$ with an activation of $\chi = 8$ [169], for which the approximation made in the second step of Eq. (3.30) can be safely applied. This leaves only the dissociation rate k_{off} open, which is supposed to convey the salt dependence of the association reaction.

Therefore, we now have all parameters at hand in order to calculate the flux to the target \tilde{j} as given by Eq. (3.29) and from this the mean target search τ time via Eq. (3.9). However, we first consider the resulting target association rate k_a . As mentioned in the very beginning of this chapter both quantities are simply related via Eq. (3.1).

3.2.1 Result obtained with the reference parameter values

The reason for starting the discussion of our results with k_a is that this is the quantity that is usually plotted in experimental papers, see e.g. Fig. 2 in [59], and thus enables us to compare our model to experimental data. Using the above mentioned parameters we obtain the central result of the present chapter: in Fig. 3.3 the target association rate is plotted as a function of the non-specific dissociation rate k_{off} . Its features, in particular the non-monotonic dependence of the target detection rate on the non-specific affinity, will be discussed in much detail below.

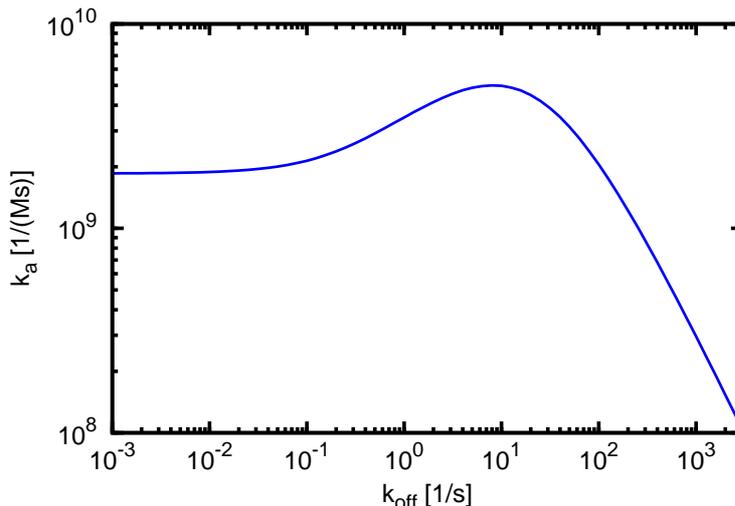


Figure 3.3: Target association rate, k_a , as a function of the dissociation rate, k_{off} . The values of all other parameters are detailed in the text. The result can be compared to the estimate of $10^8 \text{ M}^{-1}\text{s}^{-1}$ obtained with Smoluchowski's formula.

The target association rate, k_a , shows a behaviour in Fig. 3.3 which is typical for facilitated diffusion models. Namely, three different regimes can be observed. For very small dissociation rates, the target association rate is nearly independent of the exact value of k_{off} . For larger values of k_{off} , the association rate then increases and reaches a maximum. Increasing k_{off} even further leads to a dramatic decrease of the association rate. Before we explain the features of this graph, it is important to note that the maximum of the association rate, $k_a \approx 5 \cdot 10^9 \text{ M}^{-1}\text{s}^{-1}$, is relatively close to the value which was measured by Riggs and co-workers [55]. Thus, our model is able to reproduce these early experimental results fairly well.

The shape of Fig. 3.3 can be better understood when focusing on the mean search time instead of k_a , which are directly related by virtue of Eq. (3.1). The result for the mean target association time, τ can be written as

$$\tau = \tau_1 + N(\tau_2 + \tau_3) - \tau_4. \quad (3.31)$$

This is the central result of this chapter and its individual terms will be discussed below. Apart from the last term which is solely a small correction term, this result is similar to the one of Berg and Blomberg [84]. The two characteristic times τ_1 and τ_2 are even identical. New is the form of N and in particular the time scale τ_3 due to the presence of the two conformational states.

The first term in Eq. (3.31) is given by $\tau_1 = -L\dot{G}(0)$ and denotes the time which passes until the first non-specific binding to the DNA occurs²

$$\tau_1 = \frac{R_2^2}{2D_3} \left\{ \frac{1 - R_1^2/R_2^2}{k_{\text{on}}R_1/D_3} + \frac{\ln(R_2/R_1)}{1 - R_1^2/R_2^2} \right\} - \frac{R_2^2}{8D_3} \{3 - R_1^2/R_2^2\}. \quad (3.32)$$

²Note that by assumption the TF starts its search in the bulk solution and not every encounter with the DNA leads to binding as evidenced by the finite value of k_{on} appearing in Eqs. (3.32) and (3.33).

Since large values of R_2 correspond to large systems, this time increases with R_2 and becomes smaller if D_3 is large, indicating fast three-dimensional transport.

The second characteristic time is defined as $\tau_2 = -L\dot{g}(0)$ and is evaluated as

$$\tau_2 = \frac{R_2^2 - R_1^2}{2k_{\text{on}}R_1}. \quad (3.33)$$

This is the mean time that is spent in the bulk solution after a dissociation event. Accordingly, as in the case of τ_1 it becomes larger when R_2 is increased. Additionally, its value decreases when the intrinsic non-specific reaction rate k_{on} increases.

The next characteristic time scale, τ_3 , denotes the mean time that the TF spends bound to the DNA, irrespective of whether in the recognition or search mode. It is given by

$$\tau_3 = \frac{1}{\vartheta} \left(\frac{1}{k_{\text{rs}}} + \frac{1}{k_{\text{sr}}} \right). \quad (3.34)$$

Our result replaces the simpler $\tau_3 = 1/k_{\text{off}}$ of Berg and Blomberg's one-state model [84].

In Eq. (3.34), $1/k_{\text{rs}}$ and $1/k_{\text{sr}}$ denote the mean times spent in recognition and search mode before switching to the other. Thus, the sum in the bracket represents the mean for once switching back and forth between the two states. Finally, ϑ is defined as $\vartheta = k_{\text{off}}/k_{\text{sr}}$. Its inverse which appears in Eq. (3.33) determines how often the TF switches back and forth between the two conformational states before it eventually dissociates, underlining the interpretation of τ_3 .

The most interesting parameter, N , denotes the mean number of search rounds consisting of successive search in 1D and 3D. Using the abbreviations, $x_r = \sqrt{D_r/(k_{\text{rs}}L^2)}$ and $x_s = \sqrt{D_s/(k_{\text{sr}}L^2)}$, N is given by

$$N = \frac{\vartheta}{x_r} \coth\left(\frac{1}{x_r}\right) + \sum_{m=1}^{\infty} \frac{2\vartheta}{p_r(m) \left(p_r(m) \left[p_s(m) + \vartheta \left(1 - \frac{Lg_m(0)}{2} \right) \right] - 1 \right)}, \quad (3.35)$$

where $p_i(m) = 1 + m^2 \pi^2 x_i^2$, again with $i = r, s$.

Within this scheme, $N\tau_2$ denotes the total time spent *off* the chain, and accordingly $N\tau_3$ the total time spent *on* the DNA. To be more precise, these two times are only counted after the first non-specific binding event took place and finally the correction term $\tau_4 = 1/k_{\text{rs}}$ has to be subtracted. The reason for that is, that in the very last round of one-dimensional diffusion, the target will be found and thus, the last configurational change to the search state which happened in all previous rounds will not occur.

The corresponding plot of τ is depicted in Fig. 3.4, where the same data set as in the previous Fig. 3.3 was used. In this plot optimal search conditions correspond to a minimum in τ . For biologically relevant parameters, the mean search time τ is mainly determined by the two terms $N\tau_2$ and $N\tau_3$. This can be seen in Fig. 3.4, where they were plotted in addition to the mean total search time. For very low values of k_{off} , the mean search time is given to a good approximation by $N\tau_3$, and conversely, for large values of k_{off} , by $N\tau_2$.

Importantly, as is obvious from Fig. 3.4, the minimum in search time is not attained when equal amounts of time are spent in the bulk solution and on the DNA, i.e. when the two dashed curves for $N\tau_2$ and $N\tau_3$ intersect. For an optimised search, more time has to

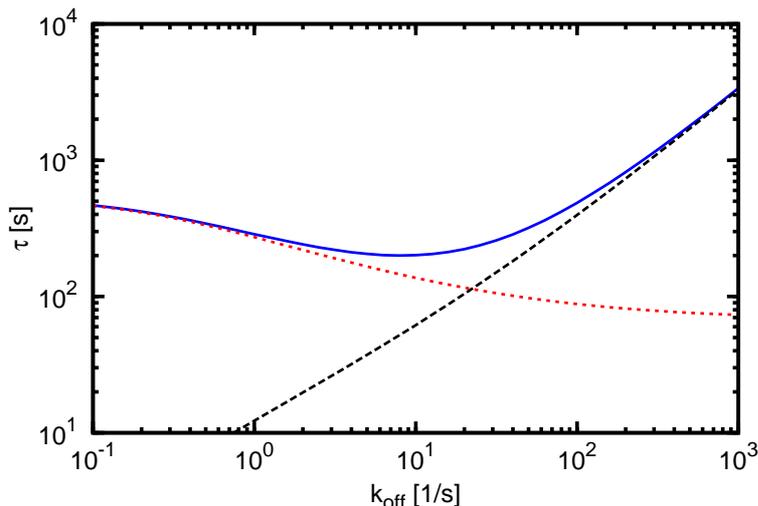


Figure 3.4: The mean target association time τ (blue line) as a function of the dissociation rate k_{off} . The dashed black line shows the contribution of the term $N\tau_2$ and the dashed red line the one of $N\tau_3$. Note that compared to Fig. 3.3 this plot shows a more restricted range of values of k_{off} .

be spent bound to the DNA, unlike in the simpler result by Slutsky and Mirny [126]. In fact, for the parameter values chosen here, the optimal fraction is $\approx 73\%$. Interestingly, after our publication on which this chapter is based was published, Petter Hammar and co-workers observed experimentally that the optimal bound fraction in a living *E. coli* cell at room temperature is 70% [32, 174].

With the above interpretations at hand, we are now able to explain the non-monotonic feature of Fig. 3.3, and thus also of Fig. 3.4. For very small values of k_{off} , in this case $k_{\text{off}} < 10^{-2}/\text{s}$, the dissociation rate is more than six orders of magnitude smaller than the switching rate from search to recognition state. Thus, it can be safely assumed that the protein never leaves the chain after the first non-specific association. Consequently, in this situation the exact value of k_{off} does not matter. The search time is then simply the sum of the time of first non-specific association and the time it takes to slide to the target. This will be shown formally in section 3.4. For biologically relevant parameters τ_1 can become negligible, such that this is an effectively one-dimensional search. However, this is inefficient since it leads to oversampling, i.e. the particle often returns to stretches of DNA which it already probed for the target, compare Pólya's early results [78].

Considering high dissociation rates, $k_{\text{off}} > 10^2/\text{s}$, this implies that the DNA chain is left rapidly. Now, the assumption that the target can only be found by sliding impacts the association rate negatively. At fast dissociation rates, the TF simply spends too few time on DNA to probe for the target, an effect that becomes worse and worse, if the residence time on DNA becomes shorter. In other words, the TF spends too much time in 3D, where the target cannot be found.

However, in between these two extreme regimes, the target search can be optimised. For the parameters chosen in Fig. 3.3, at $k_{\text{off}} \approx 10/\text{s}$, the particle stays long enough on

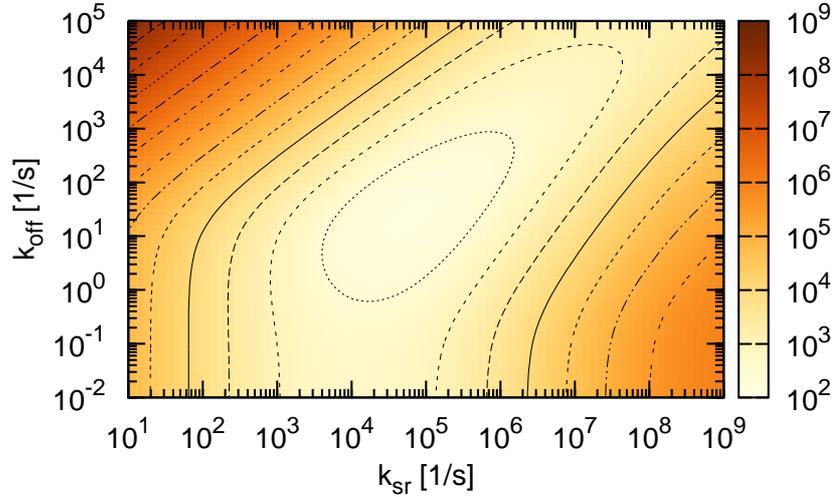


Figure 3.5: Contour plot of the mean search time as a function of the dissociation rate k_{off} and the conversion rate k_{sr} .

the DNA to look for the target, but leaves it soon enough before the one-dimensional search becomes too redundant. Note that the peak is more pronounced when choosing smaller values of the activation χ , i.e. when diffusion in the recognition mode is not much slower than in the search state (compare Fig. 4 in [174]).

In the following subsection, we study how the mean target search time depends on various parameters. Apart from the parameters which are varied in each individual figure, the remaining parameters have the values as specified in our reference set. We start the discussion with the conformational switching rates.

3.2.2 Dependence on the switching rates k_{rs} and k_{sr}

As mentioned before the rates of conversion between search and recognition state are not known experimentally. Thus, in the following we vary them in a wide range. Fig. 3.5 shows a contour plot of the mean target search time as a function of the dissociation rate k_{off} and the conversion rate k_{sr} . At the same time, the other interconversion rate is fixed to the same value as before: $k_{\text{rs}} = 10^5/\text{s}$.

Here and in the following, bright regions correspond to short search times, while dark regions correspond to an inefficient search³. First of all, it is apparent from Fig. 3.5 that there is a global minimum of the search time. Thus, for otherwise fixed parameters it is possible to tune the switching rate to the target-sensitive recognition mode and the dissociation rate from the DNA, such that the search is optimised.

³Note that throughout this work a color map has been used that is specifically designed for the needs of colour-blind people. It was obtained from www.sron.nl/~pault/gnuplot/gnuplot-ref-colmap.plt and is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported License.

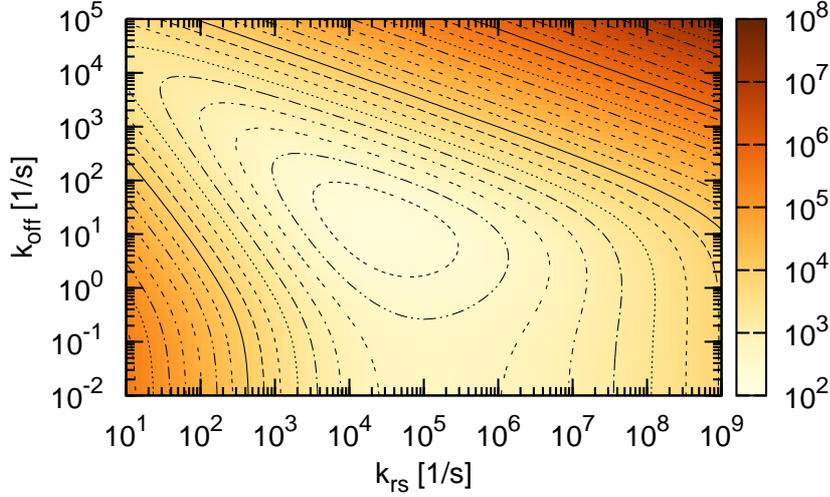


Figure 3.6: Contour plot of the mean search time as a function of the dissociation rate k_{off} and the conversion rate k_{rs} .

Furthermore, when switching to the recognition mode is very slow, $k_{\text{sr}} < 10^2/\text{s}$, increasing the dissociation rate always leads to a slower search. Thus, when this switching is slow, as much time as possible should be spent on the DNA, since only rarely the TF probes for the target. In general, as was already observed in Fig. 3.3, for small values of k_{off} , the search time becomes essentially independent of the dissociation rate. This is indicated by the (nearly) vertical shape of the contour lines in the lower part of Fig. 3.5.

If, however, switching to the recognition mode occurs at a sufficiently high rate, there is always an optimal residence time on DNA which minimises the search time. The corresponding optimal dissociation rate increases with increasing values of k_{sr} as shown by the orientation of the “valley” towards the upper right corner in Fig. 3.5.

Fig. 3.6 shows the complementary situation when the switching rate k_{sr} is fixed to a value of $10^4/\text{s}$ and k_{rs} is varied. Besides, the dissociation rate is varied as in Fig. 3.5. Again, we observe that the target search can be globally optimised. Furthermore, there exist two qualitatively differing regimes: For very large switching rates, $k_{\text{rs}} \approx 10^9$, the mean search time increases monotonically with k_{off} . This means that when the TF returns quickly to the search state, at least it should not dissociate too soon from DNA to ensure an effective probing for the target. For lower values of k_{rs} , there is again an optimal dissociation rate, which increases for decreasing values of k_{rs} . This is manifested by the valley in Fig. 3.6 which points to the upper left corner.

The observation that the target search can be optimised, when either one of the two conformational switching rates is fixed and the other is varied, raises the question if there is a global minimum when both rates are variable and only the dissociation rate k_{off} is fixed. This situation is studied in Fig. 3.7, where we chose $k_{\text{off}} = 10/\text{s}$, close to the minima in the two previous figures.

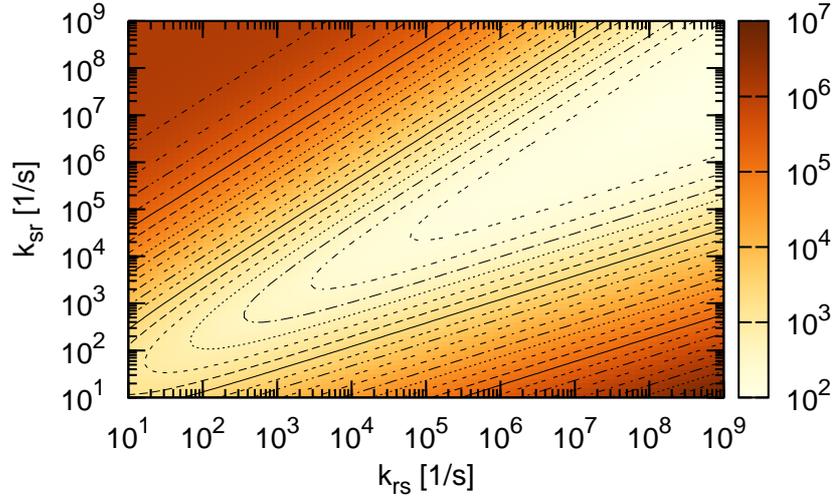


Figure 3.7: Contour plot of the mean search time as a function of both conversion rates k_{sr} and k_{rs} for a fixed value of $k_{off} = 10/s$.

An inspection of Fig. 3.7 shows that the search time can be optimised whenever any one of the two switching rates is fixed. But the absolute target search time becomes smaller and smaller, when k_{sr} and k_{rs} are increased simultaneously, while keeping them in a suitable relation. Here fast transitions between the two conformations make sure that the system is equilibrated. Then the maximum amount of probability density is supplied to the target.

In the following, we fix the switching rates k_{rs} and k_{sr} again to their reference values and study the dependence of the target search time on the various diffusion coefficients.

3.2.3 Dependence on the diffusion coefficients

Dependence on the three-dimensional diffusion coefficient Fig. 3.8 depicts the mean target search time as a function of k_{off} and the three-dimensional diffusion coefficient D_3 . This dependence is of importance, especially if one wishes to describe the *in vivo* situation later on, since the crowded interior of a living cell is expected to slow down the three-dimensional diffusion of macromolecules considerably.

Deliberately, we also consider the non-physical situation when D_3 is smaller than the one-dimensional diffusion coefficient in the search state, $D_s = 5 \times 10^{-14} \text{ m}^2/\text{s}$. In this range, it is observed that increasing the dissociation rate always increases the mean search time. This is no surprise, since changing to a slower transport mode which is blind for the target cannot accelerate the search process.

Whenever 3D diffusion is considerably faster than 1D diffusion, there is a dissociation rate which optimises the search process. Interestingly, the optimal residence time in this regime only increases a mere factor of three when D_3 is varied over nearly five orders of

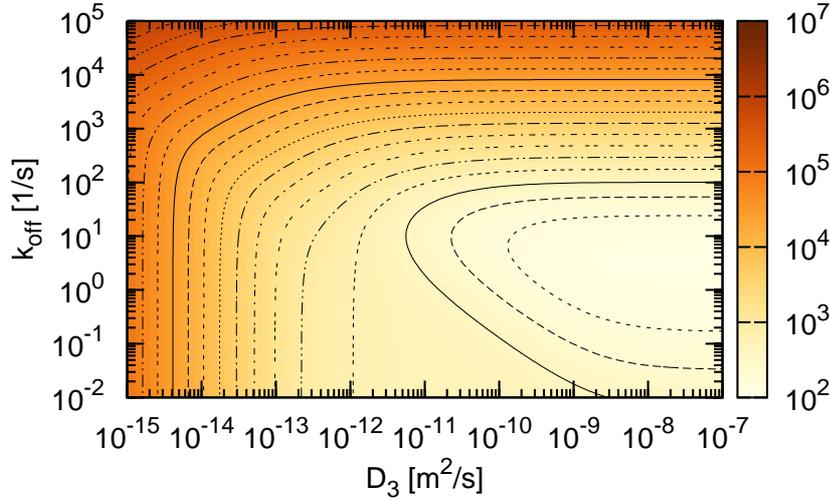


Figure 3.8: Contour plot of the mean search time as a function of the dissociation rate k_{off} and the three-dimensional diffusion coefficient D_3 .

magnitude. To understand this behaviour, it has to be stated that the beneficial influence of the 3D search phase stems from the fact that it re-equilibrates the position along the DNA cylinder after a dissociation event. If now diffusion in 3D is fast enough such that the next binding event occurs at an essentially random position, there is not much to be won by further increasing D_3 . This is similar to the effect observed by Reingruber and Holcman [168]: having a searching particle which does *not* switch between conformations, soon there will be a less than average probability density near the target, simply because it already ended up in the target. However, a searcher which can switch to a faster, but target-insensitive state, will have a much more homogeneous probability distribution. If this particle switches back to the target-sensitive state, its probability density near the target will be much higher [168]. Thus, for an efficient target detection it is important that the probability distribution is homogenised when it switches back to the target sensitive mode, but the search is not sped up if this happened even earlier. Of course, in the present case the search state itself is not sensitive for the target. However, in comparison to the unbound state, a particle in the search mode is one step closer to target detection. Thus, the same line of reasoning applies.

Dependence on the one-dimensional diffusion coefficients Now, we consider the variation of the mean search time with respect to changes in the one-dimensional diffusion coefficient D_s . In Fig. 3.9, the three-dimensional diffusivity, D_3 is again fixed to its reference value of $50 \frac{(\mu\text{m})^2}{\text{s}}$ and the two one-dimensional diffusion coefficients are varied.

This plot shows less interesting features than the previous ones. The search time decreases monotonically with the two one-dimensional diffusivities. For very low values of D_r , the search time is essentially independent of the rapidity of motion in the search

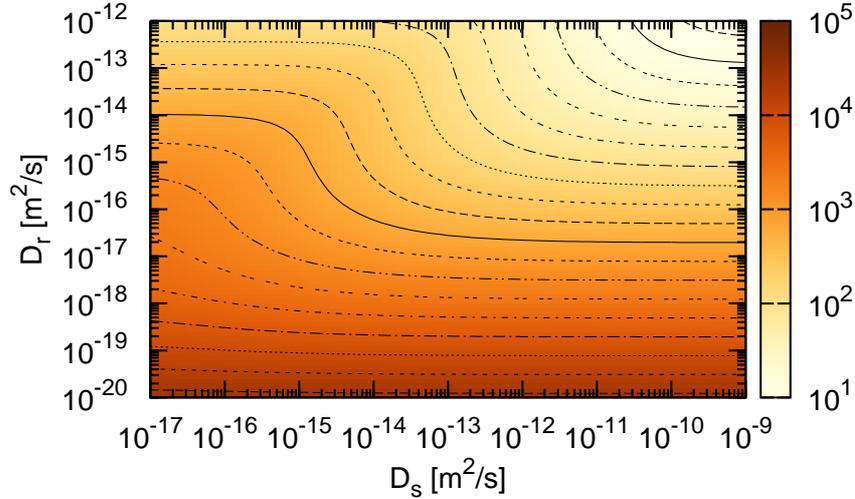


Figure 3.9: Contour plot of the mean search time as a function of the one-dimensional diffusion coefficients D_s and D_r .

state. However, for sufficiently large values of D_r there is a threshold value of D_s at which the mean search time jumps from one regime in which it is nearly independent of D_s to another one in which this is again the case, but with a reduced search time. Importantly, we do *not* see an overshoot effect as in Hu et al. [167]. The reason is simply that within our model the target detection probability in the recognition mode amounts to 100%. Thus, increasing D_r always helps to speed up the process.

In Fig. 3.10 k_{off} is fixed again to 10/s and the two diffusion coefficients D_s and D_3 are varied. Simultaneously with D_s , the diffusion coefficient in the recognition mode, D_r is varied while keeping the above mentioned ratio $D_r/D_s = \exp(-8)$.

When all other parameters are fixed, the mean search time decreases when the three-dimensional diffusion coefficient, D_3 , and/or the one-dimensional coefficient in the search state, D_s , are increased. This is not surprising, since high values of these two parameters imply a fast transport in both dimensionalities. However, the limitations of decreasing the target search time by increasing these diffusion coefficients will be discussed after describing the features of the following plot.

Both in Fig. 3.9 and in Fig. 3.10 we saw that increasing D_s leads to a reduced search time. However, as noticed by Schraner and Richter for physical reasons the one-dimensional diffusion coefficient cannot be chosen arbitrarily high when the dissociation constant is supposed to be small at the same time [83].

3.2.4 Dependence on the reaction volume

Finally, we study how the mean target search time depends on the reaction volume. For this purpose, in Fig. 3.11 the mean search time is plotted as a function of the radius R_2

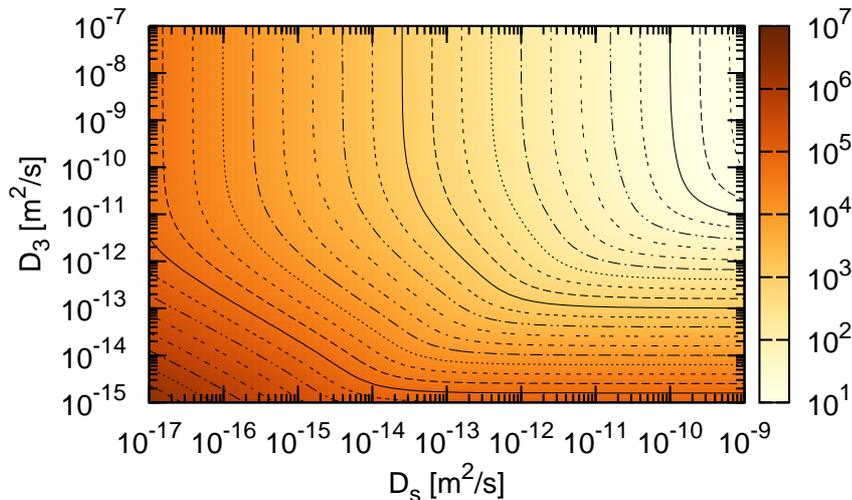


Figure 3.10: Contour plot of the mean search time as a function of the one-dimensional diffusion coefficient D_s and its three-dimensional analogue D_3 .

and the dissociation rate k_{off} .

Apparently, for small dissociation rates, the search time does not change much with variations in the value of R_2 . This is clear, since in the case of no or just a few dissociation events, it does not matter how large the three-dimensional volume actually is. Conversely, when dissociations are much more frequent, i.e. for large values of k_{off} the target search time increases steeply when the reaction volume grows.

The results presented so far all described the situation of straight DNA segments. For this situation we were able to find closed-form solutions for the two kernel functions F and G which contain all the information about the DNA conformation. In the following section, we study the situation of coiled DNA.

3.3 Results for coiled DNA

In the case of coiled DNA, it is assumed that a particle which dissociates from the DNA immediately loses its memory of the dissociation position. Recalling the form of Eq. (3.15), this implies, $g_m(u) = 0 \quad \forall m \geq 1$. This facilitates the calculations of the relevant quantities considerably as detailed below. However, we first note that the function G describing the “virgin flux” remains unaffected, since in the derivation of this function for straight DNA it was assumed anyway that the particle starts its motion homogeneously distributed in the bulk solution, such that it binds to a random position on DNA.

With the knowledge that nearly all functions g_m vanish, the mean number of search rounds to be performed can be determined via Eq. (3.35). Introducing the auxiliary

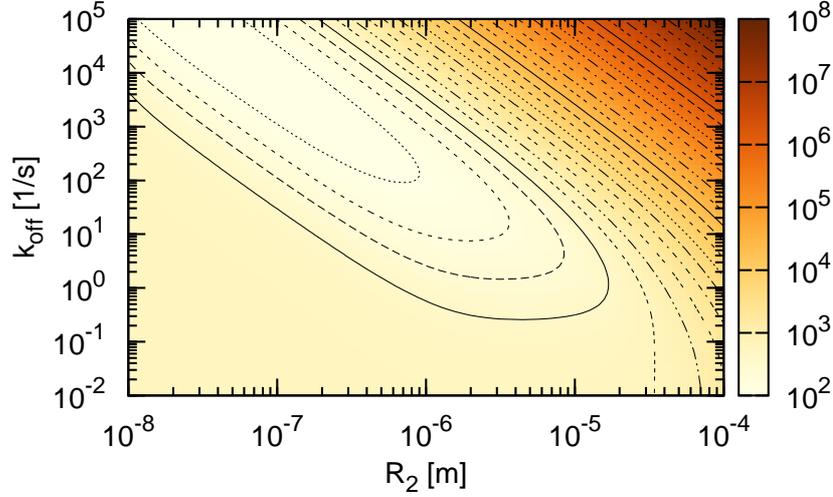


Figure 3.11: Contour plot of the mean search time as a function of radius R_2 which scales with the size of the search volume and the dissociation rate k_{off} .

functions, X and y_{\pm} defined as

$$X = \sqrt{x_s^4 - 2x_r^2x_s^2(\vartheta - 1) + x_r^4(1 + \vartheta)^2} \quad \text{and} \quad y_{\pm} = \frac{x_r^2(1 + \vartheta) + x_s^2 \pm X}{2x_r^2x_s^2}, \quad (3.36)$$

one obtains

$$N = \vartheta \left(1 + \frac{X + (x_s^2 - x_r^2(1 + \vartheta))}{2x_r^2X} f(y_+) + \frac{X - (x_s^2 - x_r^2(1 + \vartheta))}{2x_r^2X} f(y_-) \right). \quad (3.37)$$

Since in this limiting case our result is nearly equivalent to the one obtained by Jürgen Reingruber and David Holcman, we used their notation $f(x) = \coth(\sqrt{x})/\sqrt{x} - 1/x$ [168, 169]. All relations to other models will be discussed in the following subsection 3.4, where in particular we will show that this approximation corresponds to an extreme case of coiled DNA.

Since by design the mean time after which the TF returns to the DNA, τ_2 , is equal for straight DNA and for coiled DNA, it becomes obvious that the distribution of TFs equilibrates much faster when the DNA is coiled. This is beneficial for the search process, as becomes apparent in Fig. 3.12, where we plot the ratio of the mean target search time obtained with the straight DNA conformation with the one obtained for coiled DNA. The parameters used in Fig. 3.12 are as in Fig. 3.5. Bright regions correspond to parameter values which yield similar search times, while the darker regions signify an acceleration due to the coiled conformation of DNA. Most importantly we note that in a wide range of parameters the conformation does not play a large role, as becomes obvious through the white spaces in Fig. 3.12.

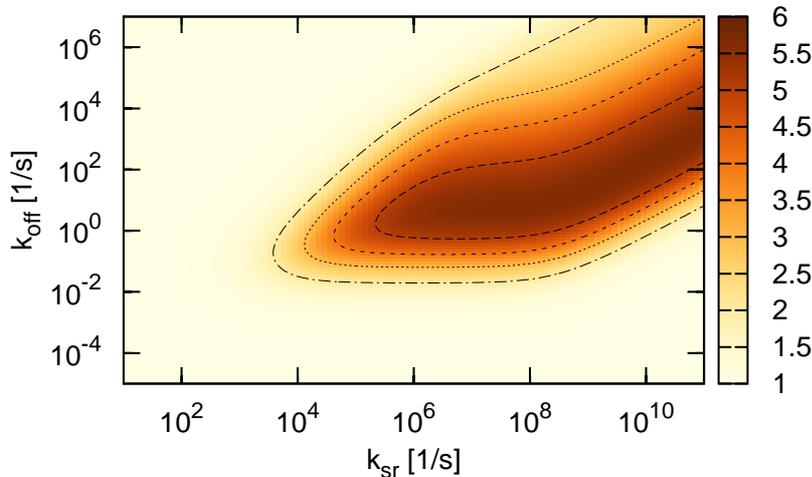


Figure 3.12: Acceleration of the target search process when a coiled DNA conformation is assumed instead of straight DNA segments. The ratio of the search time with straight DNA segments with the corresponding times for coiled DNA, $\tau_{\text{straight}}/\tau_{\text{coil}}$ is plotted. The parameters are as in Fig. 3.5. In a large range of parameter values there is very little acceleration in the coiled conformation.

In particular, when the dissociation rate k_{off} and/or the conversion rate k_{sr} is small the times are similar. In the first case, this is obvious: if the particle rarely or never leaves the chain the search time cannot depend on the conformation, since we effectively deal with a one-dimensional system. The particle associates to a random position on DNA and then slides towards the target.

In the second case, when the recognition state is only rarely entered, the probability density equilibrates very fast in comparison to this slow time scale such that again the exact conformation is not important. A similar argumentation holds for very high dissociation rates, where the times also become similar. If the particle spends most of the time in the bulk solution, the distribution equilibrates quickly, rendering the conformation unimportant again.

In the biologically relevant case of intermediate rates, the faster equilibration in system with coiled DNA can lead to target detection times which are at most a factor of approximately 6 shorter. This can be compared to the experimental results for the protein *EcoRV*, where also a rather a modest yet measurable speed-up was noticed for a coiled versus a straight DNA configuration [137, 138].

3.4 Relation to previously published models

Since the approach presented in this chapter was designed as an extension of the classical model introduced by Berg and Blomberg, we expect our model to reduce to the latter in

some limiting case [84]. Besides, as already mentioned the model also converges to the result obtained by Jürgen Reingruber and David Holcman [168, 169]. We start with the latter case.

Relation to the model of Reingruber and Holcman Eqs. (3.36) and (3.37) shown above constitute the limiting case of our model which corresponds to coiled DNA. To be able to compare our model to the approach of Reingruber and Holcman one has to adjust the initial condition to the one they chose. Adapting the initial condition to the situation where initially the particles are distributed homogeneously in the recognition mode ($n_r(z, t = 0) = n_r(t = 0) = 1/L$) and where there are no particles in the bulk solution ($G(z, t) = 0$), one obtains a result which coincides with the one published by Reingruber and Holcman [169]. In our notation this reads:

$$\tau = (N - \vartheta)(\tau_2 + \tau_3). \quad (3.38)$$

In fact, the result Reingruber and Holcman published one year earlier, can be obtained by setting $k_{\text{off}} = 0$ [168]:

$$\tau = \left(\frac{1}{k_{\text{rs}}} + \frac{1}{k_{\text{sr}}} \right) \frac{x_r^2 + 3x_s^2 f(1/x_r^2 + 1/x_s^2)}{3x_r^2(x_r^2 + x_s^2)}. \quad (3.39)$$

This corresponds to a one-dimensional search process, in which dissociation of the particle is disabled.

Relation to the model of Berg and Blomberg To study the convergence of our model the one of Berg and Blomberg [84], we study the limit of very fast switching rates between the two conformations and/or long chains. Technically this implies: $D_r \ll k_{\text{rs}}L^2$, and $D_s \ll k_{\text{sr}}L^2$. In this limit, the correction term $\tau_4 = 1/k_{\text{rs}}$ can be safely neglected and the number of search rounds as given by Eq. (3.35) is well approximated by

$$N \approx \sum_{m=1}^{\infty} \frac{2\vartheta}{\vartheta(1 - Lg_m(0)/2) + m^2\pi^2(x_r^2 + x_s^2)}. \quad (3.40)$$

Blurring the dynamical differences between search and recognition mode by setting $D_1 = D_s = D_r$ and $k_{\text{rs}} = k_{\text{sr}}$, we obtain that the mean time spent on the chain is simply $\tau_3 = 2/k_{\text{off}}$ and that N simplifies to

$$N \approx \sum_{m=1}^{\infty} \frac{2k_{\text{off}}}{k_{\text{off}}(1 - Lg_m(0)/2) + 2m^2\pi^2D_1^2/L^2}. \quad (3.41)$$

This agrees with the classical result by Berg and Blomberg if their one-state dissociation rate k_{off} is identified with the present $k_{\text{off}}/2$ [84]. At first sight, the factor of two seems to manifest a difference, but it is only due to our assumption $k_{\text{rs}} = k_{\text{sr}}$ which implies that—neglecting the effect of the target—equal amounts of time are spent in both conformational states. Thus, to have the same physical picture as in the classical one-state model, the dissociation rate which only acts on particles in the search state has to be two times higher.

Of course, the approximations for coiled DNA can also be applied to the original one-state model of Berg and Blomberg [84]. With the two characteristic times τ_1 and τ_2 remaining unaffected, one obtains again

$$\tau = \tau_1 + N(\tau_2 + \tau_3), \quad (3.42)$$

where τ_3 is now simply given by $1/k_{\text{off}}$ and

$$N = \sqrt{\frac{k_{\text{off}} L^2}{D_1}} \coth \left(\sqrt{\frac{k_{\text{off}} L^2}{D_1}} \right) - 1 = \frac{k_{\text{off}} L^2}{D_1} f \left(\frac{k_{\text{off}} L^2}{D_1} \right). \quad (3.43)$$

Once more in the second identity we used the notation of Reingruber and Holcman [168]. Apparently, this has the same form as Eq. (2.20) and the similar result of Coppey et al. [127]. The result coincides completely with Berg's original result, when the macroscopic dissociation rate in Eq. (2.20) is equal to the microscopic one, or in other words when every microscopic dissociation leads to a complete loss of correlation. This is an extreme case of coiled DNA and in accordance with our interpretation of the disappearance of nearly all coefficients g_m .

Using an expansion for small arguments x , namely $x \coth(x) - 1 \approx x^2/3$, one obtains $\lim_{k_{\text{off}} \rightarrow 0} N = 0$. But since τ_2 is inversely proportional to k_{off} , the mean search time τ in the one-state model as given by Eq. (3.42) results in this limiting case as:

$$\lim_{k_{\text{off}} \rightarrow 0} \tau = \tau_1 + \frac{L^2}{3D_1}. \quad (3.44)$$

This result⁴ has a straightforward interpretation. The mean search time in the absence of dissociation is simply the sum of the mean time it takes for the first non-specific association, τ_1 , and the mean search time in a one-dimensional interval of length L . In fact, the second term is equal to Eq. (2.10) which was derived using Szabo's first passage time formalism [68].

We are now able to compare the results of our full model with three limiting cases which correspond to previously published models. These are the original one-state model by Berg and Blomberg and its limiting case for coiled DNA as well as the same limiting case for our two-state model. Identical parameters were taken for all four cases, but apparently in the case of the one-state models, the two conformational switching rates are no longer present and there is a single one-dimensional diffusion coefficient, D_1 . For the latter we use a value of $D_1 = 0.0455 \frac{(\mu\text{m})^2}{\text{s}}$ —close to the experimental value reported by Elf and co-workers [39]—such that $D_1 = D_a$ with the apparent diffusion coefficient D_a defined via Eq. (3.30).

In Fig. 3.13 the target association rates obtained with these four models are plotted as a function of the dissociation rate k_{off} . The blue lines show the results of the two-state model, while the dark red lines the corresponding ones of the one-state model. In both cases, full lines are for straight DNA segments, and dot-dashed lines for coiled DNA.

⁴In fact, the result of Reingruber and Holcman, Eq. (3.39), generalises Eq. (3.44) to the two-state model [169].

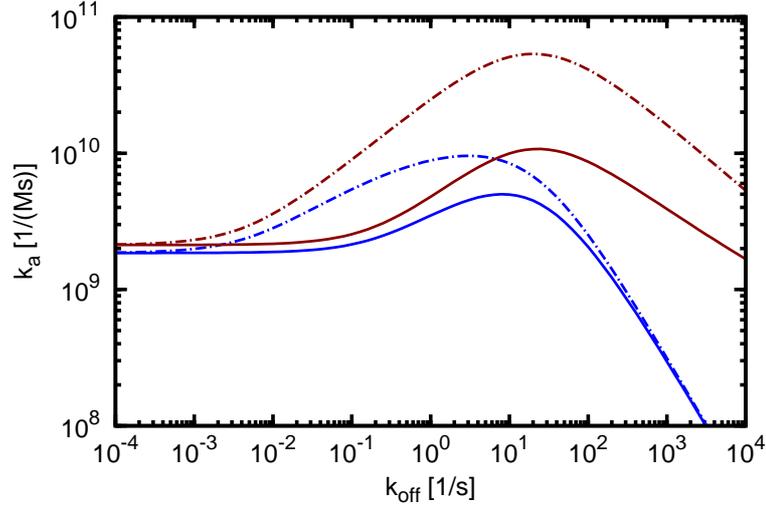


Figure 3.13: The target association rate k_a in four models as a function of the dissociation rate k_{off} . Blue lines refer to the two-state model, while red lines represent the one-state model. In both cases full lines correspond to a straight DNA conformation and dot-dashed lines to coiled DNA.

The most important observation is, that in all four cases the three typical regimes are observed. Besides, as already mentioned in section 3.3 for very high and very low dissociation rates, the target association rate in the two-state model becomes independent of the underlying DNA conformation. This can be seen from the convergence of the blue lines in both extreme regimes. As evidenced by the red lines in Fig. 3.13, this is also true for the one-state model. In fact, for $k_{\text{off}} \rightarrow 0$ the corresponding mean search time is given by Eq. (3.44).

A closer inspection of Fig. 3.13 shows that in this limit, although we chose the one-dimensional diffusion coefficients such that both effective diffusivities are the same, the target association rates in both models are *not* the same. In the one-state model target association happens faster, since every target encounter leads to detection, which is not the case in the two-state model. In the opposite limit of high dissociation rates, all target detection rates decrease. However, in the case of the two-state models this decrease is more severe ($k_a \propto k_{\text{off}}^{-1}$), than for the one-state model, where $k_a \propto k_{\text{off}}^{-1/2}$. The reason for this is that in the two-state model after re-association the particle is initially in the target-insensitive search mode, which is left quickly in this regime. Conversely, in the one-state model the full bound time can be used to probe for the target. This different scaling behaviour can in principle be used to distinguish experimentally between both types of models.

Again in accordance with the findings of section 3.3 the target is more quickly found on coiled DNA, since equilibration is obtained faster. Apart from this general observation, one notices that the dissociation rate which optimises the search is smaller in the case of coiled DNA. Since the equilibration is more efficient, less time in the bulk is needed and more time can be spent sliding.

Finally, target search in the two-state model is always slower than in the corresponding one-state model. This does not contradict the result of Reingruber and Holcman who proved that introducing a second state in which a particle diffuses faster decreases the search time, even if the particle is not able to detect the target in this new state [168]. In the present case, the comparison is between two states whose effective one-dimensional diffusivities are the same, but the particle in the one-state model has better chances to actually detect the target.

3.5 Summary and outlook

In this chapter it was shown that it is possible to generalise the classical facilitated diffusion model of Otto G. Berg and co-workers to the case when two conformations of the searching particle are explicitly taken into account. Introduced by other authors in order to resolve the *speed-stability paradox*, these two conformations differ in the way binding to the DNA is concerned. While it is in the “search” state, the particle can quickly slide along DNA, but it cannot detect the target. Conversely, in the “recognition” mode sliding motion is very slow, but the particle is able to detect the target. Importantly, this model assumes *a priori* that the target is detected by sliding in the recognition mode and not by direct detection from the bulk solution.

Just like in the original one-state model one observes the typical three regimes of facilitated diffusion models and accordingly our generalised approach converges to these previously published models in appropriately chosen limiting cases. In particular the target detection rate is a non-monotonic function of the non-specific dissociation rate from DNA. This represents a trade-off between spending too much time with redundant one-dimensional search or spending too much time in the bulk solution where by definition the target cannot be found.

Unlike in simplified versions of the model, we can rationalise that it is beneficial for the repressor to spend more than half of the search time on DNA. The exact value of the optimal partition is, of course, dependent on the choice of system parameters, but for the values chosen in Fig. 3.3 our result is close to the experimental result by Hammar and co-workers [32].

A weak spot of the model is that switching between the two states occurs stochastically. Such blind switching is definitely not an efficient means to detect the target as is evidenced by reduced target detection rates with respect to the corresponding one-state models (compare Fig. 3.13). However, our two-state model is able to satisfy simultaneously the complementary demands of tight binding to the target and fast sliding along the DNA, while the classical one-state models assume that this can be achieved with a single state. The extension step to the case when a switch of conformation is coupled to the underlying nucleotide sequence will be undertaken in chapter 5.

Within our blind switching approach it was observed that the search can be optimised by appropriately choosing one interconversion rate, if the other one has a fixed value. As we assume that target detection occurs with 100% efficiency in the recognition mode, there is no “overshooting” effect and the detection rate increases with the diffusivity in the recognition mode. The same applies for the diffusion coefficients in the search mode and in 3D since they convey faster transport. In the same way smaller system sizes via a

reduced value of the outer cylinder's radius are beneficial. Finally, when comparing the search for a target on straight DNA segments to an extreme model of coiled DNA, it was observed that in the latter case search can be moderately sped up in a certain range of parameters.

The assumption of a reflecting boundary condition at the outer cylinder has already been criticised in 1977, since it introduces correlations along the cylinder [86]. But in the dilute *in vitro* situation that this model aims to describe the outer boundary is so far away from the DNA chain that this effect is not important. This is different in the crowded interior of a living cell. The following chapter is devoted to the question how the search of a TF in such an environment can be modelled, when neighbouring DNA segments are way closer.

4 In vivo facilitated diffusion model

In the following chapter we are still concerned with the generic problem of the association of proteins with a specific target on DNA. If necessary, we will use parameter values for the lac system in *E. coli*. The model presented in the last chapter and most of the other published models aim at describing the situation in an *in vitro* experiment. But apparently, this can only be an intermediate step, because in the long run one wishes to explain the behaviour in living cells. To do this, we will partition the cell volume and especially the genome it contains in structural subunits. In each of these we will solve analytically a microscopic facilitated diffusion model. These solutions will be used as an input for the numerical description of the search process in the whole living cell. This hybrid semi-analytical approach bridges the gap between purely theoretical models which can only be solved in simple geometries on the one side and purely numerical approaches on the other side.

In general, when trying to describe the search problem in a living cell, one has to answer two fundamental questions:

1. What is different in living cells? *and*
2. Is it possible to translate or rather transcribe previous models to account for these changes?

To answer the first question, we note that unlike in controlled *in vitro* experiments, a living cell contains many other molecules. This *macromolecular crowding* was already mentioned in chapter 1 and influences chemical reactions in cells in many ways. For example, it shifts equilibria of reactions towards the associated state and importantly for this study, the diffusion of proteins in such crowded solutions will be slowed down. The question whether or not the resulting motion of macromolecules is still Brownian or anomalous, is highly debated, since there is evidence for both interpretations [39,93,94]. For the moment, we will assume herein that the motion of the lac repressor in a living *E. coli* cell is still Brownian, but characterised by a reduced diffusion coefficient¹. The effect of anomalous transport mechanisms on the search efficiency will be briefly discussed towards the end of this chapter. Since we saw in the previous chapter that a reduced diffusion coefficient in 3D slows down the motion, this is a negative effect in terms of the search efficiency. But as pointed out by Tabaka et al. more crowding implies that the DNA inside the cell is confined to a smaller volume such that the search volume is reduced [142]. This is expected to reduce the search time. Roughly speaking, these crowding effects concern the three-dimensional part of the search process in a cell.

However, also the one-dimensional search phase will be affected, since the particles which bring about the crowding effects, the so-called crowding agents, will bind to the

¹The general question how the diffusional behaviour of particles changes in a structured environment was studied by the author of this thesis in two other works [180,181].

DNA. This has two main consequences: the amount of base pairs to which the searching protein can bind is reduced and since it is usually assumed that such a bound protein cannot be by-passed by the searcher, the sliding motion is hindered [122,140–142]. We will approach these issues mainly in the next chapter 5, while the present one focuses on another important difference to the dilute *in vitro* situation: the conformation of DNA in a bacterial cell. We will try to shed light on this in the following section, before section 4.2 aims at answering the second of the above questions. Namely, it describes how we adapt the facilitated diffusion model to describe the *in vivo* situation.

This chapter reviews and extends the results which have been published previously in [182].

4.1 Organisation of bacterial DNA

There are at least two fundamental differences between the conformation of DNA that is studied in *in vitro* experiments and the one that is encountered in living cells. First, they are confined to the cell body, which is much smaller than the volume that a relaxed DNA coil would require. This can be understood by noting that the contour length of the *E. coli* genome consisting of approximately 4.6 million base pairs, is $4.6 \times 10^6 \times 3.4 \times 10^{-10} \text{ m} \approx 1.6 \text{ mm}$, while the longest axis of a bacterial cell is in the micrometer range. Thus, the genome is packaged in the cell volume. Even more severely, although we were right to characterise prokaryotes as living organisms that do not have a cell nucleus in the introductory chapter 1, they do have a similar object, the *nucleoid*². Unlike nuclei in eukaryotes, nucleoids in prokaryotes do not have a nuclear membrane. The second difference with respect to DNA in *in vitro* experiments is that several proteins act together in the compaction of bacterial DNA, giving it an internal structure.

Let us summarise what is currently known about the structural organisation of bacterial DNA in living cells [183]. The bacterial DNA making up the genome can be thought of as existing in connected subunits, which we call *blobs*. They can change their form in time. By means of atomic force microscopy, their size in *E. coli* cells was estimated to be in the range of several tens of nanometers [184]. Other experiments yielded evidence for subunits containing several tens of kilobasepairs which have a diameter of $(70 \pm 20) \text{ nm}$ [185, 186]. These data sets will be used as an quantitative input for our semi-analytical model.

Complementary results obtained with the bacterium *Caulobacter crescentus*, whose genome spans a similar amount of base pairs like *E. coli*, are even more relevant for the design of our approach. Using carbon copy chromosome conformation capture, commonly denoted as *5C*, several research groups obtained a three-dimensional model of *Caulobacter*'s genome [187]. In fact, *5C* is an extension of *3C* which is short for chromosome conformation capture. These results further substantiated earlier findings that the position of genes in real space along the long axis of the cell correlates strongly with the position of the corresponding gene on the chromosome map [188].

The growing amount of experimental knowledge, of course, found the interest of scientists aiming to model bacterial genomes. Some studied the organisation under the premise that regulatory interactions are responsible for the positioning of genetic

²We note that for some species the nucleoid can fill the entire cell volume.

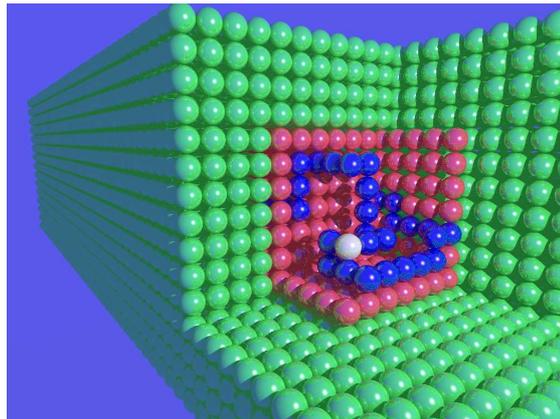


Figure 4.1: Scheme of the model genome. The red spheres show the nucleoid volume, while the green spheres represent the rest of the cell volume. Within the nucleoid, blue spheres denote blobs which contain (non-target) DNA, and the single white sphere is the target blob.

loci [117, 118]. Other researchers modelled bacterial DNA as a ring polymer and focused on the role of entropy [189, 190]. In the following we adapt the simple, yet convincing model proposed by Mathias Buenemann and Peter Lenz [191]. Their geometric model is based on a self-avoiding walk (SAW) description of the structural subunits of DNA. It will be presented in more detail in the following subsection.

4.1.1 Model genome

The bacterial genome is modelled as a sequence of spheres, in the following also called “blobs”, which reside on a lattice. The spheres represent the structural subunits mentioned before and the lattice represents the volume of the nucleoid³. To mimic the rod-like shape of the bacterial cell, the underlying lattice is assumed to have two short axes of equal length and one longer axis. The genome is modelled as a closed SAW, and a small piece of it can be seen as the blue (and white) spheres in the schematic Fig. 4.1.

Herein, the green spheres represent the volume of the cell which does not contain DNA, while the red spheres stand for the nucleoid volume. Blue spheres denote the blobs within the nucleoid, which actually contain DNA, but not the target DNA. The single white sphere represents what is called the target blob in the following. It is simply the blob which contains—alongside many non-specific base pairs—the target sequence. Note that Fig. 4.1 is a slice through the real situation, in which spheres in the foreground would block the view on the nucleoid blobs and particularly the target blob.

The “pearl necklace” of DNA-containing blobs (shown in blue in Fig. 4.1) is in this work obtained similarly to the original treatment by Buenemann and Lenz [191]: one starts with a closed loop oriented along the cell’s long axis which starts at the bottom of the cuboid, reaches the top and subsequently returns to the bottom. Initially, such a loop with minimal contour length is chosen and then the necklace is extended by inserting

³This constitutes a difference to the treatment of Buenemann and Lenz who took the whole cell volume [191]. But as mentioned before for some species the nucleoid fills the whole cell volume.

hooks of minimal length at random positions. This continues until the predefined contour length is reached. Throughout this procedure elongation steps which leave the volume of the nucleoid are rejected. These preliminary conformations are subsequently equilibrated via transformations taken from the algorithm introduced by Madras, Orlicsky and Shepp (MOS) [192]. Only transforms which do not destroy the SAW nature of the necklace are performed and counted are only transformations after which the genome still resides within the nucleoid volume.

The search space is finalised by putting the nucleoid volume and the DNA conformation it contains in the centre of the larger cell volume. Even though the DNA conformation may vary dynamically in a real cell, we keep this conformation fixed throughout the target search simulation. This latter assumption is motivated by recent findings that the influence of DNA dynamics on the search process is not too large [136].

In a simplistic view, we observe that for the target search process it only matters if a lattice site is occupied by a blob or if it is empty. Before we detail how the target search is modelled, we emphasise that two parameters were introduced within this model: first, the radius of gyration of one blob, r_g , which fixes the lattice spacing to twice this value, $d = 2r_g$ and second, N_b which denotes the number of base pairs which are found in one blob. Both will attain values which we take from experimental studies.

4.2 General search model

We now present our general model for the target search of a TF in a living cell, which we subsequently call *in vivo facilitated diffusion model* (IVFDM). As our previous GFDM, it describes the search process of a single searching particle and again we assume that the target lies in the middle of one “blob” in our model genome. To be more precise, a blob in the middle of the pearl necklace is designated to be the target blob, such that it will be typically close to one of the ends along the longest axis of the nucleoid. Before we describe the details of our model, it is appropriate to ask if one cannot directly extend the previous model to the *in vivo* situation. At least such a straightforward extension was possible in other approaches [122, 141].

Direct extension of GFDM? In principle, one can rescale the lengths, diffusion coefficients and reaction rates which enter the GFDM to approach the situation in a living cell. This straightforward method, however, has some limitations. First of all, it is highly questionable to approximate the DNA in a living cell as a straight DNA or a random coil. Besides, even if this approximation can be applied, when one plugs in typical values for the nucleoid volume, one obtains that the outer effective cylinder radius of GFDM, R_2 , which denotes a typical distance between vicinal DNA segments, becomes smaller than the DNA-protein contact radius, R_1 . In this case, GFDM can no longer be applied as becomes obvious e.g. in Eqs. (3.32) and (3.33).

4.2.1 Details of the search model

In a rough scheme, the target search of the TF is a random walk on the lattice of spheres as shown in Fig. 4.1 which was described in the previous section. However, it is not a

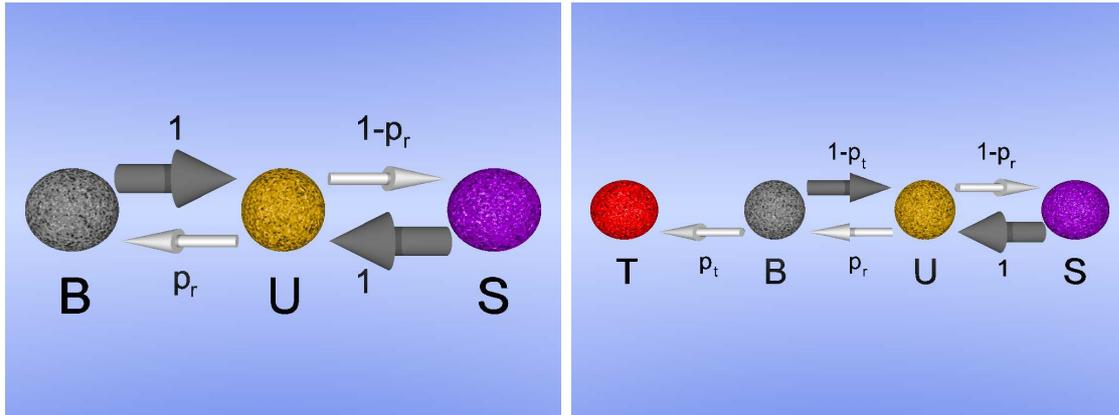


Figure 4.2: Scheme of the search process within a non-target blob (left panel) and within the target blob (right panel). B represents TFs which are bound to the DNA inside the blob, U denotes TFs which are unbound, but within the blob and S those who are searching outside the blob. In the right panel which describes the situation in the target blob, the new state T represents a particle which has detected the target. That is why not all particles return to state B , as implied by the grey arrow from B to U which is thinner as the grey arrow from S to U (compare also the explicitly given probabilities).

simple random walk in the sense that not all lattice sites are equivalent. Namely, there are three types of sites: empty sites, sites with a blob but without the target and finally a single blob containing the target. The two latter types are exclusively found within the nucleoid volume, even though this volume also contains empty sites. In the remaining cell volume there are only empty sites. Importantly, unlike the DNA the searching particle can be located outside the nucleoid. Thus, the cell volume is roughly divided in two parts: the inner part with the nucleoid, containing DNA, and an outer part void of DNA.

As in GFDM the process starts unbound, at a random position in the volume. Neglecting the presence of the target in a first theoretical step, the state of the searching particle can be classified into three groups (see the left panel of Fig. 4.2).

If it is on a lattice site without DNA, it has no other options than being unbound. Since it is *searching* for some stretch of DNA, we denote this state by S . Because it performs a random walk in a confined volume, with certainty, i.e. probability one, at some point it will encounter a blob which contains DNA. This is represented in the left panel of Fig. 4.2 by the thick dark grey arrow⁴ down to the right. As in a simple random walk, a single step starting from an empty lattice site, takes a time span, $\tau_{3D} = (2r_g)^2/(6D_3)$. Since the particle will enter a blob with DNA in the *unbound* state, we denote this state by the letter U .

Once inside this blob, there are two possible outcomes: either the particle binds to the

⁴The thickness of an arrow in Fig. 4.2 is proportional to the probability with which the corresponding event occurs. The corresponding probability additionally appears next to an arrow. While in the left panel both arrows pointing towards state U have the same size, this is no longer the case in the right panel.

DNA inside the blob, or it leaves the blob without attaching to DNA. In the following section we will present a microscopic model for these mutually exclusive processes, but for the moment we will just denote the probability that a first binding event occurs by p_r (compare the left panel of Fig. 4.2). This *bound* state is represented by the letter B . As we assumed that the blob does not contain the target, with certainty (again probability 1) after some time the particle will dissociate and return to the unbound state inside the blob, U . Then the particle faces the same question as before: will it bind again (with probability p_r) or will it leave the blob (with probability $1 - p_r$). Thus, typically the particle will undergo several rounds of binding and unbinding before it eventually leaves the blob. Obviously, the number and duration of such rounds will be determined by the microscopic parameters.

So far, we excluded the blob containing the target from our considerations. Inside this blob the scheme of states as depicted in the left panel of Fig. 4.2 has to be extended. The states S , U and B are still present. However, when the TF binds to the DNA inside the target blob, now it is possible that it detects the target before it dissociates. This successful event is shown in the right panel of Fig. 4.2, where the new state T was introduced.

As in GFDM we assume *a priori* that the target can only be found via sliding, or in the parlance of the present model: state T is only entered from the bound state B . This happens with a probability p_t to be detailed in subsection 4.3.1. With the complementary probability $(1 - p_t)$ the particle will dissociate from DNA before detecting the target (compare the right panel of Fig. 4.2). Accordingly, also in the target blob typically several rounds of binding and unbinding occur, before either the target is found or the particle leaves the target blob. Motivated by the set-up of many experiments, the simulation stops when the target is detected, and we do not consider what happens after an eventual dissociation from the target. We now discuss how to obtain the probabilities p_r and p_t and the corresponding time scales in terms of a microscopic model which is based on the general facilitated diffusion scheme.

4.3 Microscopic model

The microscopic model describes the situation within a DNA-containing blob. We start its description with the simpler derivation of the target detection probability, p_t , before we detail the calculation of the probability to attach to DNA, p_r in subsection 4.3.2.

4.3.1 Derivation of the target detection probability, p_t

Inside a blob we invoke the usual assumption that the DNA of contour length $2L$ forms a random coil and that the target is located exactly in the middle of the chain. Then, the first binding position on DNA is completely random and the site where the particle dissociated and where it re-associated are practically uncorrelated, compare [127] and section 3.3. Some computational studies focused on more detailed local structures, e.g. rosettes and studied different positions of the target within a rosette [162]. For our present semi-analytical model we invoke a more coarse-grained picture.

Again denoting the one-dimensional diffusion coefficient by D_1 and the microscopic dissociation rate by k_{off} , the following simple differential equation for the density of

bound TFs at time t and position z , $c(z, t)$, holds:

$$\frac{\partial c(z, t)}{\partial t} = D_1 \frac{\partial^2 c(z, t)}{\partial z^2} - k_{\text{off}} c(z, t). \quad (4.1)$$

Note that this corresponds to a one-state model. In principle, it could be extended to a two-state model but we chose to keep the number of free parameters as low as possible.

The assumption of a random first binding position translates to $c(z, t = 0) = 1/L$. The boundary conditions $c(z = 0, t) = 0$ and $\left. \frac{\partial c(z, t)}{\partial z} \right|_{z=L} = 0$ assume a perfect target detection on encounter and that the end of the DNA is reflecting [127]. The latter choice is on the one hand not important, since the DNA inside a blob is rather long, and on the other hand motivated by the observation that the compaction of the genome has to be brought about by some structural proteins which cannot be by-passed. Importantly, for the search process it implies that one cannot slide to a neighbouring blob. Apart from that the DNA is assumed to be void of other bound particles. While this is certainly not true in a living cell for the biologically relevant short sliding lengths this assumption is not too wrong [104]. This is in agreement with the theoretical prediction that *in vivo* sliding will not play a role on length scales of more than “a few tens of base pairs” [140]. The consequences of the presence of other non-specific binders will be studied in more detail in chapter 5.

Again, Eq. (4.1) is most conveniently solved in Laplace space with respect to time, yielding:

$$\tilde{c}(u, z) = \frac{1}{L(u + k_{\text{off}})} \left(1 - \frac{\cosh((L - z)\sqrt{\frac{u + k_{\text{off}}}{D_1}})}{\cosh(L\sqrt{\frac{u + k_{\text{off}}}{D_1}})} \right). \quad (4.2)$$

As usual u denotes the variable complementary to time. The flux of particles into the target, $\dot{j}_{\text{target}}(t) = D_1 \left. \frac{\partial c(z, t)}{\partial z} \right|_{z=0}$, then follows in Laplace space as:

$$\tilde{j}_{\text{target}}(u) \simeq \frac{\tanh(L/\ell)}{L/\ell} + \frac{u}{2k_{\text{off}}} \left(\frac{1}{\cosh^2(L/\ell)} - \frac{\tanh(L/\ell)}{L/\ell} \right), \quad (4.3)$$

where the Taylor series in u was expanded up to linear order and again $\ell = \sqrt{D_1/k_{\text{off}}}$.

While in the previously studied cases the 0th order term was always equated to be 1, for the first term in Eq. (4.3) this is in general not the case. The term linear in u will be discussed below. In fact, the first term is the first quantity we wanted to derive: the probability to detect the target before dissociating, p_t :

$$p_t = \frac{\tanh(L/\ell)}{L/\ell}. \quad (4.4)$$

This probability solely depends on the ratio between the half-length of DNA within a blob, L , and the typical sliding length, ℓ . Fig. 4.3 shows p_t as a function of this ratio.

Since for small arguments L/ℓ , $\tanh(L/\ell) \approx L/\ell$, p_t approaches unity if the sliding length is much larger than the DNA length. This is clear, since in this case the total length of DNA will be explored before dissociation and combined with an infinitely fast target detection on encounter, one can be sure to detect the target. In the opposite limit

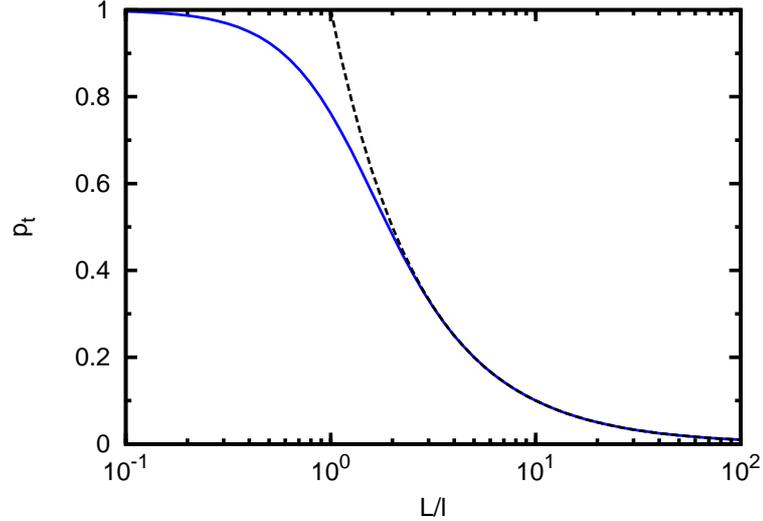


Figure 4.3: The target detection probability p_t (blue line) as a function of the ratio L/ℓ . For long DNA chains, p_t is well approximated by $1/(L/\ell)$ (dashed black line).

of large DNA lengths, one has $\tanh(L/\ell) \approx 1$, such that $p_t \approx \ell/L$. This limit has been drawn as a dashed black line in Fig. 4.3. Thus, when the DNA is sufficiently long, the probability to detect the target converges to the ratio of the typical sliding length and the DNA half-length. In between these two extreme regimes p_t decreases monotonically as a function of L/ℓ .

Up to a factor of u/p_t , the second term in Eq. (4.3) is the conditional target detection time τ_t . Here, conditional detection time means the mean target detection time given that the target is detected at all. One obtains:

$$\tau_t = \frac{1 - 1/(p_t \cosh^2(\frac{L}{\ell}))}{2k_{\text{off}}} = \frac{1 - \frac{\ell}{L} / (\sinh(\frac{L}{\ell}) \cosh(\frac{L}{\ell}))}{2k_{\text{off}}}. \quad (4.5)$$

Conversely, the probability to dissociate before the target is detected is given by $1 - p_t$. This can be obtained by calculating the Laplace transform of the global dissociation flux which is defined as $k_{\text{off}} \int_{z=0}^L \tilde{c}(u, z) dz$. The associated conditional dissociation time τ_d is given by

$$\tau_d = \frac{1}{2k_{\text{off}}} \left(3 - \frac{\tanh^2(L/\ell)}{1 - \tanh(L/\ell)/(L/\ell)} \right). \quad (4.6)$$

In blobs which do not contain the target, the DNA is simply left after an average time span of $1/k_{\text{off}}$. This can also be seen by letting $L \rightarrow \infty$ in Eq. (4.6). In reference [182] a differing convention was applied: namely, for τ_d the time span $1/k_{\text{off}}$ was always used instead of Eq. (4.6). However, we checked that this only slightly modifies the results for biologically relevant values. In particular, the qualitative conclusions drawn in reference [182] are still true.

Thus, we derived the “one-dimensional” probabilities and their corresponding time scales. But before we can detail the calculation of the mean target search time in section 4.4, we first have to calculate the DNA association probability, p_r and its associated time scales.

4.3.2 Derivation of the non-specific association probability, p_r

In order to derive p_r , i.e. the probability to bind to DNA instead of leaving the blob via diffusing out with diffusivity D_3 , we again have to solve a diffusion equation. This time, however, in three dimensions and in a spherical domain. More explicitly, we study two concentric spheres: an inner one with radius r_g which corresponds to the radius of gyration of the blob and an outer one with radius r_2 . The latter radius marks a cut-off, i.e. when a particle reaches this radial distance, it is assumed that it “forgot” where it came from and continues its search in a neighbouring blob. For the ease of notation we introduce the dimensionless parameter α via $r_2 = \alpha \cdot r_g$, where $\alpha > 1$, but otherwise *a priori* arbitrary. We will detail below which value of α we will use for the numerical evaluation. Within the inner sphere of radius r_g there is a homogeneous distribution of DNA to which the particle can bind.

The following derivation of p_r is similar to a calculation in [86]. Since we assume spherical symmetry, the functions solely depend on $r = |\mathbf{r}|$ and not on the full vector \mathbf{r} . Introducing the non-specific association rate k_{ass} per base pair, which is measured in units of $\text{M}^{-1}\text{s}^{-1}$, we have to solve the following differential equation for the three-dimensional density of particles, $c(r, t)$:

$$\frac{\partial c(r, t)}{\partial t} = \begin{cases} D_3 \Delta c(r, t) - \kappa c(r, t), & 0 < r < r_g \\ D_3 \Delta c(r, t), & r_g < r < r_2 \end{cases}, \quad (4.7)$$

where Δ denotes the Laplace operator and $\kappa = nk_{\text{ass}}N_b$. As before N_b denotes the number of base pairs within a blob and n stands for the density of DNA such that κ has the physical dimension of 1/time unit.

Since we assume that after dissociation the particle immediately loses track of its point of dissociation and that the same is true for a particle entering the blob from outside, this corresponds to the following initial condition:

$$c(r, t = 0) = \begin{cases} c_0, & 0 < r < r_g \\ 0, & r_g < r < r_2 \end{cases}. \quad (4.8)$$

As we study the case of a single searching particle in a blob, we use $c_0 = n = 3/(4\pi r_g^3)$, where n denotes the density of DNA within a blob. Finally, we use the boundary condition, $c(r = r_2, t) = 0$, to take into account that particles which reach this cut-off distance leave the blob and thus the microscopic system.

Again going to Laplace space with respect to time and exploiting the spherical symmetry of the system, the differential equation (4.7) becomes:

$$u\tilde{c}(u, r) = \begin{cases} c_0 + D_3 \Delta \tilde{c}(r, u) - \kappa \tilde{c}(r, u), & 0 < r < r_g \\ D_3 \Delta \tilde{c}(r, u), & r_g < r < r_2 \end{cases}. \quad (4.9)$$

There are two possible ways for the particle to leave the three-dimensional domain of the system:

1. it can dissociate away by reaching the radial distance $r = r_2$. In Laplace space, the corresponding flux, $\tilde{j}_{\text{out}}(u)$, is given by

$$\tilde{j}_{\text{out}}(u) = -4\pi r_2^2 D_3 \left. \frac{\partial \tilde{c}(u, r)}{\partial r} \right|_{r=r_2}, \quad \text{or} \quad (4.10)$$

2. within the inner sphere, i.e. within the blob, it can bind non-specifically to the DNA. In Laplace space, this flux, $\tilde{j}_{\text{bind}}(u)$, is given by

$$\tilde{j}_{\text{bind}}(u) = 4\pi\kappa \int_0^{r_g} dr r^2 \tilde{c}(u, r). \quad (4.11)$$

Plugging in the solution of Eq. (4.9), one obtains

$$\tilde{j}_{\text{out}}(u) = \frac{3}{r_g^3 q_1^3} \frac{r_2 q_2}{\sinh(q_2 \delta r)} \frac{q_1 r_g \coth(q_1 r_g) - 1}{\coth(q_1 r_g) + \frac{q_2}{q_1} \coth(q_2 \delta r)}, \quad \text{and} \quad (4.12)$$

$$\tilde{j}_{\text{bind}}(u) = \frac{3}{r_g^3 q_1^3} \frac{\kappa}{u + \kappa} \left[\frac{r_g^3 q_1^3}{3} - \frac{(q_1 r_g \coth(q_1 r_g) - 1)(1 + r_g q_2 \coth(q_2 \delta r))}{\coth(q_1 r_g) + \frac{q_2}{q_1} \coth(q_2 \delta r)} \right], \quad (4.13)$$

where we introduced the notation $q_1 = \sqrt{\frac{u+\kappa}{D_3}}$, $q_2 = \sqrt{u/D_3}$ and $\delta r = r_2 - r_g$. Due to the relations

$$\tilde{j}_{\text{bind}}(u) \simeq p_r(1 - \tau_b u), \quad \text{and} \quad \tilde{j}_{\text{out}}(u) \simeq (1 - p_r)(1 - \tau_e u), \quad (4.14)$$

expanding the Maclaurin series up to linear order yields after a lengthy but straightforward calculation for the association probability:

$$p_r = 1 - \frac{3\alpha\phi(\gamma)}{\alpha + (\alpha - 1)\gamma^2\phi(\gamma)}. \quad (4.15)$$

In Eq. (4.15) we used the auxiliary function $\phi(\gamma) = (\gamma \coth(\gamma) - 1)/\gamma^2$ in analogy to the function f introduced by Reingruber and Holcman and on page 62 in the previous chapter [168]. In the present case, this function involves the dimensionless quantity $\gamma = r_g/\sqrt{D_3/\kappa}$. The term in the denominator, $\sqrt{D_3/\kappa}$, has the physical dimension of a length and corresponds roughly to the distance the particle diffuses in 3D within the blob before it is captured by DNA. Importantly, p_r solely depends on γ and on the fixed parameter α .

Fig. 4.4 shows p_r as a function of γ for the dimensionless parameter $\alpha = \sqrt{23/5}$ which denotes the system's cut-off distance. As mentioned above, α is a parameter which can in principle attain any value larger than 1. However, as we will detail in a specific paragraph below in the numerical evaluation of the mean search time, we choose the above value.

An expansion for small arguments γ , yields $\lim_{\gamma \rightarrow 0} \phi(\gamma) = 1/3$. Thus, $\lim_{\gamma \rightarrow 0} p_r(\gamma) = 0$ as expected since this limit corresponds to a vanishing association rate k_{ass} . For increasing values of γ , Fig. 4.4 shows that the non-specific association probability grows monotonically. In the limit of large values of γ , one obtains $\lim_{\gamma \rightarrow \infty} p_r(\gamma) = 1$. This is coherent since

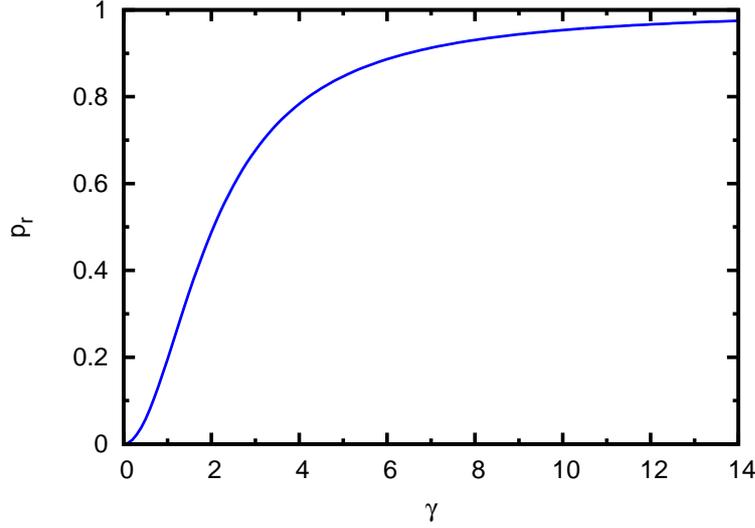


Figure 4.4: Association probability p_r as a function of the dimensionless parameter γ for $\alpha = \sqrt{23/5}$.

large values of γ correspond to fast non-specific association. However, this limit is also reached for vanishing D_3 where the interpretation is not that straightforward. Obviously, an immobile TF cannot leave a blob (as implied by $1 - p_r = 0$), but if such a particle is able to bind to DNA is a priori unclear. For the finite values of γ studied herein, it does not make a difference.

In the general case, the conditional mean time for non-specific binding is given by the following rather lengthy expression:

$$\tau_b = \frac{\alpha}{2\kappa} \left\{ 5\alpha + (4\gamma^2(\alpha - 1) - 15\alpha) \phi(\gamma) + (12 - 15\alpha + 2\gamma^2(1 - \alpha)^2) \gamma^2 \phi^2(\gamma) \right\} \\ \times (\alpha + (\alpha - 1)\gamma^2 \phi(\gamma))^{-1} \times (\alpha + (\gamma^2(\alpha - 1) - 3\alpha)\phi(\gamma))^{-1}. \quad (4.16)$$

Conversely, the conditional mean time it takes for a TF to leave the environment of a blob is given by:

$$\tau_e = \frac{1}{2\kappa} \left\{ \alpha(3 - \phi^{-1}(\gamma)) + \gamma^2 \left((3\alpha - 2)\phi(\gamma) + \frac{2 + \alpha}{3}(1 - \alpha)^2 \right) - \frac{\gamma^4}{3}(1 - \alpha)^3 \phi(\gamma) \right\} \\ \times (\alpha + (\alpha - 1)\gamma^2 \phi(\gamma))^{-1}. \quad (4.17)$$

Derivation of $\alpha = \sqrt{23/5}$ We now comment on the choice $\alpha = \sqrt{23/5}$ in our numerical evaluations. As already mentioned, the only hard restriction on α is that it should be larger than unity: $\alpha > 1$. Thus, we decided to choose the convention that in the limit where the TF has no affinity at all for DNA, $\kappa \rightarrow 0$, the escape time as given by Eq. (4.17) should be equal to the time a random walk step takes in a region void of DNA, τ_{3D} . Formally, this implies

$$\lim_{\kappa \rightarrow 0} \tau_e(\kappa) = \frac{r_g^2}{30D_3} (5\alpha^2 - 3) = \frac{4r_g^2}{6D_3} = \tau_{3D}. \quad (4.18)$$

Solving for α , one obtains: $\alpha = \sqrt{23/5} \approx 2.14$. Importantly, we note that this value was simply taken in order to obtain a meaningful scaling behaviour, but it should not be taken literally. Comparing with Fig. 4.1, one notices that for such a value of α neighbouring blob regions would overlap. Thus, its value is not dogmatic, but as we will see below it leads to a very reasonable agreement with experimental data. We now derive an explicit formula for the mean search time within our model.

4.4 Derivation of the mean target search time

Since the particle starts its search at a random position in the cell, the total search time can be divided into three parts:

1. the time that passes until the target blob is encountered for the first time, *then*
2. since not every encounter with this blob leads to a successful target detection, the second part comprises the product of how often the target blob is encountered in vain and the time it takes to return to the target blob after a missed chance to detect it *and finally*
3. the time it takes to actually detect the target in the final stage of the search process.

In all three phases we can act on the assumption that irrespective of whether or not a blob that is encountered en route contains the target, usually several rounds of binding and subsequent unbinding events occur. In fact, a blob without target can be considered to be a target-containing blob, in which, however, the target detection probability is zero, see Eq. (4.19) below.

In the absence of a target, a single round of binding and unbinding on average takes $\tau_c = \tau_b + 1/k_{\text{off}}$, where τ_b is given by Eq. (4.16). As mentioned previously, in the presence of the target, the term $1/k_{\text{off}}$ has to be replaced by τ_d as given by Eq. (4.6). In both cases the number of rounds or loops involving binding and unbinding is given by the function

$$g(\chi) = \frac{\chi}{1 - \chi}, \quad (4.19)$$

where $\chi = p_r(1 - p_t)$ denotes the probability of binding, but not finding the target.

Without the presence of the target, one has $p_t = 0$, such that $g(\chi) = p_r/(1 - p_r)$, underlining the meaningfulness of Eq. (4.19). Since leaving the blob consumes on average τ_e as given by Eq. (4.17), in a blob without target on average the following time span is spent:

$$\tau_{\text{blob}} = \tau_e + \frac{p_r}{1 - p_r} \tau_c. \quad (4.20)$$

Here the second term is the product of the average number of search rounds and of the duration of a single round. In the presence of the target, the combined probability of not detecting it before the blob is left, is given by

$$p_{\text{uns}} = \frac{1 - p_r}{1 - \chi}. \quad (4.21)$$

Here the index “uns” alludes to the fact that this an unsuccessful event in terms of the target search. Conversely, the combined probability of successfully detecting it, is given

by $1 - p_{\text{uns}}$. Note that without target, $p_t = 0$ and therefore $\chi = p_r$, such that $p_{\text{uns}} = 1$, i.e. an unsuccessful event happens with certainty, as it should be. Successfully finding the target on average takes a time span of $\tau_{\text{suc}} = \tau_b + \tau_t + g(\chi)\tau_c$. Here the third term is the total time which is spent in unsuccessful search rounds, while the first two terms correspond to the time for binding and subsequently detecting the target. Conversely, leaving the target blob empty-handed spoiled a mean time of $\tau_{\text{uns}} = \tau_e + g(\chi)\tau_c$.

With the just mentioned microscopic probabilities and time spans at hand, just four other parameters are needed to determine the mean search time. They are obtained from the numerical simulations. Both for the initial phase of the trajectory leading to the first encounter with the target blob and for a trajectory which starts at the target blob and returns to it, the number of random walk steps has to be counted ($n_{f,3D}$ and $n_{r,3D}$) together with the number of non-target containing blobs encountered while doing this ($n_{f,\text{enc}}$ and $n_{r,\text{enc}}$). Here the index ‘‘f’’ denotes that it concerns the first part of the trajectory. Similarly, the index ‘‘r’’ describes the return to the target blob.

With this notation the mean target search time becomes:

$$\begin{aligned} \tau &= (n_{f,3D} - n_{f,\text{enc}})\tau_{3D} + n_{f,\text{enc}}\tau_{\text{blob}} \\ &+ \frac{p_{\text{uns}}}{1 - p_{\text{uns}}} (\tau_{\text{uns}} + (n_{r,3D} - n_{r,\text{enc}})\tau_{3D} + n_{r,\text{enc}}\tau_{\text{blob}}) \\ &+ \tau_{\text{suc}}. \end{aligned} \tag{4.22}$$

In this central result of this chapter, the ordering in three lines is according to the three phases mentioned at the beginning of this section. In reference [182] the term $n_{i,\text{enc}}$ was not subtracted from $n_{i,3D}$ (where in both cases $i = f, r$ representing the first encounter and return trajectories). However, this only shifts the results slightly without changing any of the conclusions.

The following section is devoted to the results for the mean search time which are obtained with Eq. (4.22).

4.5 Results of the IVFDM

In the following subsection we state our reference set of parameters that are kept unchanged throughout the simulations if not stated otherwise. As before they are motivated by the lac repressor system in *E. coli*.

4.5.1 Reference set of parameters

As mentioned we use an outer cut-off radius $r_2 = \alpha \cdot r_g$ with $\alpha = \sqrt{23/5}$. Based on experimental observations we use the following diffusion coefficients to describe the situation in a living cell: $D_1 = 0.046 \frac{(\mu\text{m})^2}{\text{s}}$ and $D_3 = 3 \frac{(\mu\text{m})^2}{\text{s}}$ [39]. Apart from the dissociation rate k_{off} and the non-specific association rate k_{ass} which both are the main variables in the following, this leaves two further microscopic parameters to be determined: the radius of gyration of a blob r_g and the number of base pairs such a blob contains. Note that both are in fact average values and the latter is related to the half-length of DNA, L , that a blob contains via $L = N_b b/2$, where $b = 0.34 \text{ nm}$ denotes the length unit base pair. For these two quantities we employ two parameter sets based on different experiments. The

Parameter set	$n_{f,3D}$	$n_{f,enc}$	$n_{r,3D}$	$n_{r,enc}$
a	31514	766.41	18689	463.48
b	2594.7	175.63	1291.9	90.848

Table 4.1: Simulation results obtained with IVFDM using the parameter sets a and b .

first one, denoted by a : $r_g = 15$ nm and $N_b = 10^4$ [185, 191]. The second one, denoted by b : $r_g = 35$ nm and $N_b = 5 \cdot 10^4$ [186]. Thus, we deal with a finer model (a) and with a coarser one (b). This also determines our lattice spacing and hence our search space. For the size of the nucleoid and the cell volume we use the experiments of Suckjoon Jun and Andrew Wright [189]: the nucleoid volume is approximately a cylinder of length $l_{nuc} = 1.39 \mu\text{m}$ and of diameter $d_{nuc} = 240$ nm. The cell volume is approximated by a larger cylinder of length $l_{cell} = 2.5 \mu\text{m}$ and of diameter $d_{cell} = 500$ nm.

In our simulation we approximate the nucleoid cylinder by a cuboid with edge lengths $l_z = l_{nuc}$ (long axis) and $l_x = l_y = \sqrt{\pi d_{nuc}^2/4}$ (short axes) and similarly for the cell volume. For the parameter set a , one then has a nucleoid lattice⁵ of size $7 \times 7 \times 46$ within the cell lattice of size $15 \times 15 \times 83$. For parameter set b , we have a lattice of size $6 \times 6 \times 36$ hosting a sublattice of size $3 \times 3 \times 20$. Because there is no exact central positioning with (sub)lattices of this size, equivalent ensembles were introduced and the results were subsequently averaged over these. Since the genome of *E. coli* has approximately 4639 kbp, the pearl necklace modelling it has a length of 464 blobs for parameter set a and a length of 92 blobs for set b ⁶.

These values enable us to calculate what fraction of lattice sites is occupied by a blob. Using the whole cell volume, one obtains $\approx 2.5\%$ with parameter set a . However, since all the blobs are in the smaller nucleoid volume, there the fraction of occupied sites is approximately 20.6%. Using parameter set b , these values become 7.1% and 51.1%. At first sight, it might be surprising that the values obtained with the two sets differ so much. However, one has to notice that while the fraction of occupied sites is higher in set b , the actual density of basepairs in a blob is smaller. Apart from round-off errors due to the discreteness of our system, the volumes and the DNA content coincide.

The simulation results for $n_{i,3D}$ and $n_{i,enc}$ are stated in table 4.1. Note that these are ensemble-averaged values, but as we will detail below results obtained with single conformations do not differ much among one another.

The data presented in table 4.1 shows that as expected parameter set a constitutes a more detailed model of DNA, as more individual steps are registered. One can directly calculate the fraction of sites occupied by a blob that a particle encounters on its trajectory. For parameter set a , one obtains 2.4% for the first search phase and 2.5% for return trajectories. In the case of parameter set b , one obtains 6.8% (first encounter) and 7.0% (return). Two points are striking: first, for both sets the measured values are close to the fraction of occupied sites in the whole cell that we calculated above. Thus, during its trajectory the particle seems to explore the whole cell volume. Or in other words, it

⁵This parameter set was used for the illustration presented in Fig. 4.1

⁶Note that by design of our algorithm, we can only study even numbers of blobs. This is however, not too much of a restriction.

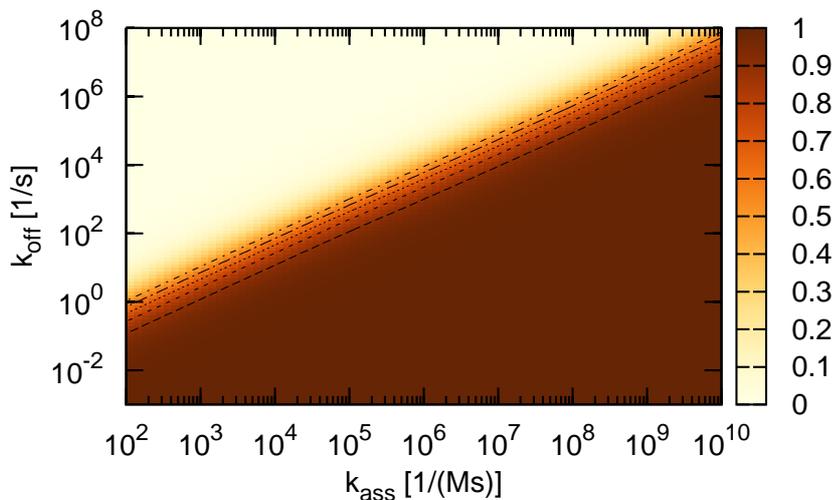


Figure 4.5: Fraction of time during which the TF is bound to DNA as a function of the non-specific association rate (k_{ass}) and the dissociation rate (k_{off}). The result was obtained with parameter set a . The contour lines are for $f_b = 0.5, 0.6, 0.7, 0.8$ and 0.9 .

diffuses through an effective medium [136]. Second, the values obtained for return trajectories are marginally larger than the ones characterising the first search phase. This can be understood as follows: while the first search phase starts at a random position, and thus with substantial probability outside the nucleoid volume, in the case of return trajectories it is known that the search starts in the nucleoid and thus in a region with an elevated DNA density.

In the next subsections we study results obtained with parameter set a . Subsection 4.5.5 will be devoted to results obtained with the alternative parameter set b .

4.5.2 Bound fraction of time

Before we actually consider the mean target search time, we first study the fraction of the total search time during which the TF is bound to the DNA, because this will ease the discussion of the following results. This value is denoted by f_b and technically it is obtained by summing up only the terms proportional to $1/k_{\text{off}}$, τ_d and τ_t in Eq. (4.22) and subsequently dividing the result by the total search time obtained with the same equation. This preparatory result obtained with parameter set a is plotted in Fig. 4.5 as a function of the dissociation rate k_{off} and the non-specific association rate k_{ass} . All other parameters are chosen as described in the previous subsection.

Bright regions in this figure correspond to trajectories in which the particle is predominantly unbound, while dark regions mark trajectories in which it is mostly bound. As expected, when considering any vertical slice through Fig. 4.5, one sees that when the dis-

f_b	c_0	c_1
0.5	1.0228	1.9495
0.6	1.0229	2.1298
0.7	1.0229	2.3258
0.8	1.0230	2.5653
0.9	1.0231	2.9258

Table 4.2: Fit parameters for the bound fraction of time, f_b as given by Eq. (4.23).

sociation rate grows, the particle spends less time bound to DNA. Conversely, when looking at a horizontal slice: increasing the non-specific association leads to a larger bound fraction. Furthermore, as is apparent from Fig. 4.5, contour lines of $f_b = 0.5, 0.6, \dots, 0.9$ are parallel, straight lines with a positive slope. More explicitly, lines of constant f_b can be fitted with the following linear equation:

$$\log_{10}(k_{\text{ass}}[1/(\text{Ms})]) = c_0 + c_1 \times \log_{10}(k_{\text{off}}[1/\text{s}]). \quad (4.23)$$

The fit values of c_0 and c_1 are presented in table 4.2. Thus, as already observed directly, that contour lines are parallel is expressed in the fact that c_0 is nearly independent of f_b . The second parameter, c_1 is an increasing function of f_b and illustrates that each contour line has a different axis intercept. Importantly, the fit values of c_0 are close to unity. This is not surprising, since the ratio of k_{ass} and k_{off} is simply the non-specific binding constant.

With this picture in mind, the Figs. 4.6, 4.8 and 4.9 below can be interpreted more easily, since they show the same range of values of k_{off} and k_{ass} and it is clear which parts of these figures correspond to parameter regimes in which the particle is predominantly bound or unbound.

4.5.3 Mean search time

We now turn the attention to the most important quantity, the mean target search time which is defined in Eq. (4.22). It is shown in Fig. 4.6 as a function of the dissociation rate k_{off} and the non-specific association rate k_{ass} . The result was obtained using parameter set a .

As usual for facilitated diffusion models, the mean search time depends non-monotonically on both rates that are varied in Fig. 4.6. This means, that in the range of rates studied here, when one of the two rates is fixed, the other one can be chosen to minimise the search time. Following the course of local minima along the descending ‘‘valley’’ towards the upper right corner of Fig. 4.6, one sees that the search time can be reduced by increasing the non-specific association rate, but at the same time the dissociation rate has to be chosen appropriately.

At first sight these local minima seem to lie on a straight line, in similarity to the lines of constant f_b seen in Fig. 4.5. If this were the case, there would be a global value of f_b that would optimise the search process. We test this hypothesis in the following subsection. There the result will also be compared to the recent experimental findings by

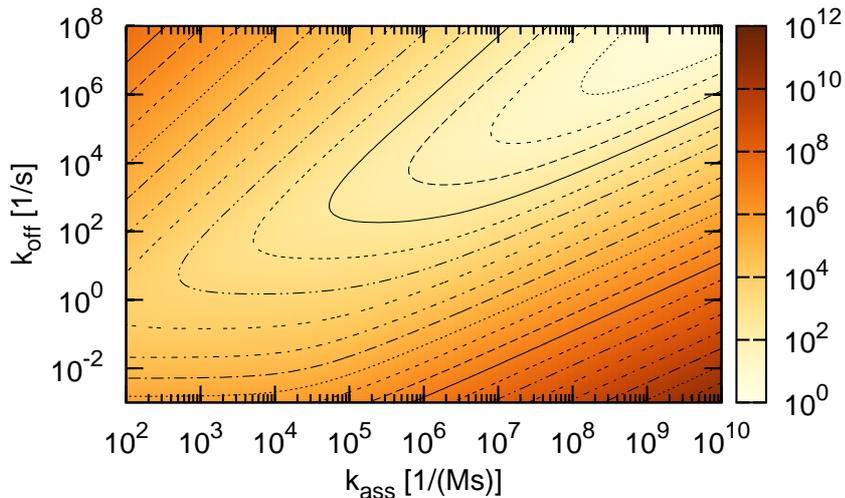


Figure 4.6: Mean target search time as a function of the dissociation rate k_{off} and the non-specific association rate k_{ass} . The results were obtained with parameter set a .

Elf and co-workers [39]. Qualitatively, our result is in agreement with the observations of Koslover et al. that the search time can be minimised when off- and on-rates are increased while keeping them in an appropriate ratio [136].

4.5.4 Searching at near optimal conditions

In the case of the lac repressor it was recently measured experimentally, that it spends approximately 87% of the time bound to DNA [39]. Of course, this value does not constitute a dogma, however the consensus both from experimental as well as from theoretical studies seems to be that the repressor spends more than half of the search time non-specifically bound to DNA. Thus, in the following we study the five equidistantly distributed values of f_b in the range $[0.5, 0.9]$ which were already highlighted as contour lines in Fig. 4.5. The mean target search times along these lines are depicted in Fig. 4.7 as a function of the dissociation rate k_{off} . Equivalently, they could have been presented depending on the association rate k_{ass} , since there is a simple linear relation between these two rates (compare Eq. (4.23)).

An inspection of Fig. 4.7 shows that for a wide range of dissociation rates, $k_{\text{off}} < 10^4/\text{s}$ the search times with different values of f_b do not differ much. We note that a similar conclusion has been drawn in a numerical study by Foffano et al. for ten proteins which search simultaneously [161]. In this regime of rather rare dissociations, the yellow curve represents the fastest search time, implying that spending a large fraction of time bound to DNA is beneficial. In turn, this implies that the association rate should be rather large. Conversely, for larger values of the dissociation rate the ordering of the curves is

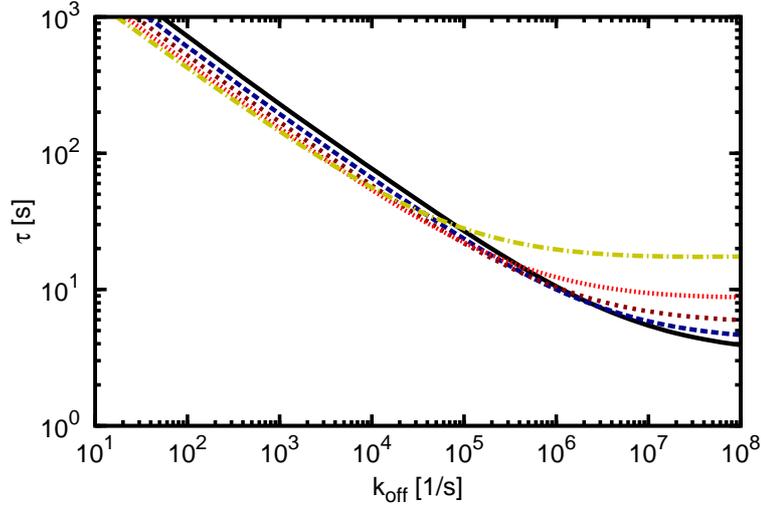


Figure 4.7: Mean target search time as a function of the dissociation rate for five different values of f_b : $\alpha = 0.5$ (black, continuous), $\alpha = 0.6$ (blue, long dashes), $\alpha = 0.7$ (dark red, short dashes), $\alpha = 0.8$ (red, dots) and $\alpha = 0.9$ (dark yellow, dot-dashed). The curves in this plot are slices through Fig. 4.6 at the corresponding contour lines. Results with parameter set a .

inverted. Thus, when dissociation occurs often, it is advantageous to spend most of the time unbound.

In general this figure shows, that within our model if one of the microscopic rates is fixed, for example due to physical or biological reasons, there is an optimal fraction of time the particle should spend on the DNA. However, there is *no* global value of f_b that optimises the search.

Within our model, at $k_{\text{off}} \approx 200/\text{s}$, the minimal search time which can be reached by choosing a non-specific association rate of $k_{\text{ass}} \approx 2.5 \times 10^5 \text{M}^{-1}\text{s}^{-1}$, is $\tau \approx 302 \text{s}$. Such a dissociation rate was discussed in Koslover et al. [136] and importantly this result is close to the experimental result of 354 s which was obtained by Elf and co-workers [39].

In general, at this value of k_{off} Fig. 4.7 shows that for the values of f_b studied here, the search times do not differ very much. A similar conclusion has been drawn by Foffano et al. in the case of ten particles which search simultaneously [161]. This nicely agrees with early experimental results which showed that unlike in *in vitro* studies, in living cells the dependence on the ion concentration and thus on the dissociation rate is not that pronounced [163].

4.5.5 Influence of different parameter sets

The results shown so far were all obtained with parameter set a . In Fig. 4.8, we now compare them to the results obtained with the alternative parameter set b , while all other parameters remained unchanged. We plot the ratio of the mean search times with set a with the ones obtained with set b as a function of k_{off} and k_{ass} .

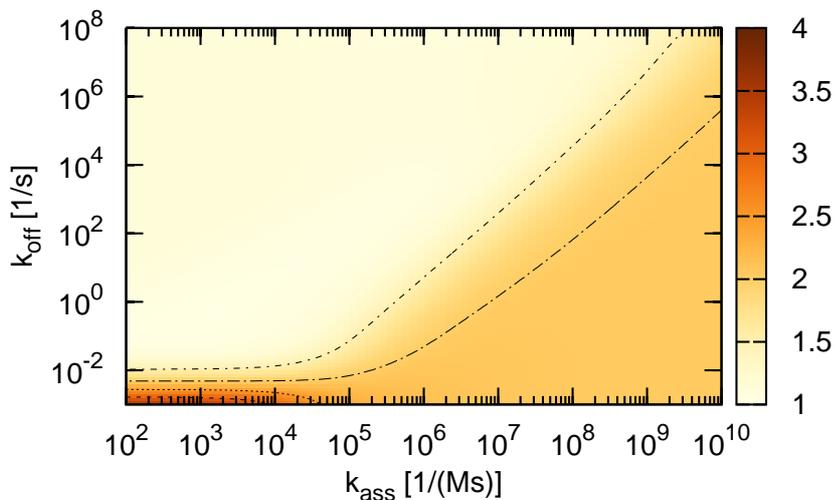


Figure 4.8: Ratio of the mean search time obtained with parameter set a with the ones obtained with set b .

In a wide range of the two parameters which are varied here, the mean search times are comparable. This is particularly true for the “interesting” regime, for example seen in Fig. 4.6, where the search time is minimised for a fixed value of the association rate. Only in the bottom left corner notable differences between the two parameter sets can be seen. In this regime, where both binding and unbinding are slow, the different amounts of base pairs a single blob contains make the difference.

Globally speaking, in terms of the mean search time, the results obtained with the two parameter sets do not differ much. One can go one step further, and ask if results using single conformations differ considerably. However, when comparing the mean search times obtained at $k_{\text{ass}} = 10^5 \text{ M}^{-1}\text{s}^{-1}$, it appears that they all are within one percent of the ensemble averaged value (data not shown), see supporting information of [182]. Thus, it seems to be fair to limit ourselves to considering the ensemble averaged values.

Likewise, it was shown by Koslover et al. in a combined theoretical and numerical study that different DNA conformations yielded similar search times [136]. Their interpretation was that within a short time span a motion arises which is characterised by an effective three-dimensional diffusion coefficient [136]. Motion is slowed down due to binding events, but the exact conformation is not too important [136]. This agrees with our observation that the medium that the particle experiences on first encounter and return trajectories has the same fraction of DNA-occupied blobs as the whole cell.

4.5.6 Acceleration due to local searches

Finally, one may relax the assumption that the search process is started at a random position in the cell. There are two reasons for doing this. First, biologically relevant is not

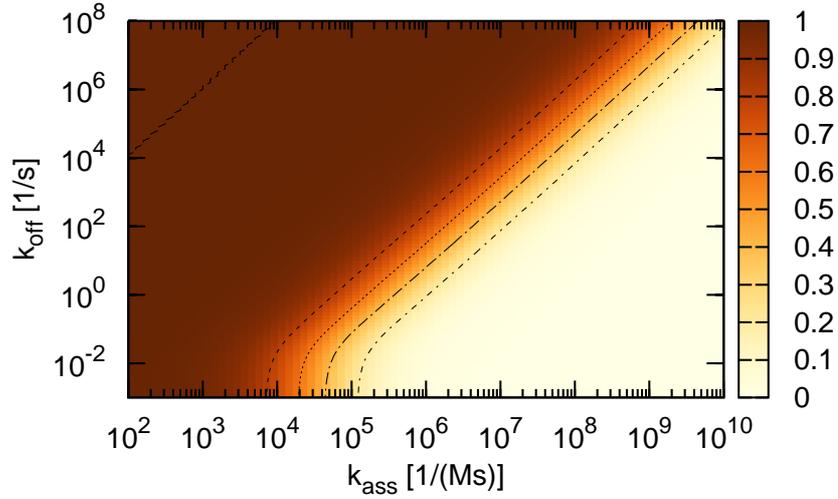


Figure 4.9: Ratio of the mean search times for a particle starting its search already in the target blob, but unbound, and for a particle starting at a random position in the cell volume. All other parameters are equal.

only the time it takes for a TF to find the operator from a random starting position, but also the time it takes to return to the target sequence after dissociating from the target⁷. Second, the co-localisation effect which was mentioned in subsection 2.4.1 [114, 115]: many TFs are expressed in the vicinity of where they later on have to perform their tasks. For example, the gene *lacI* encoding the lac repressor is near the lac promoter and its genes. It seems obvious, that a search process starting closer to the target should be faster than an equivalent one whose starting position is more distant. However, this is not true when (a) the dimensionality of the system is larger than two *and* (b) the starting position of the searcher is not too close to the target [74]. In this case the search time is independent of the initial separation [74]. Thus, as in Pólya's classical result, 2 is the critical dimension. Note that distance effects in gene regulation were also studied in [120].

In Fig. 4.9 we study the influence of the starting position on the search times within our model. This figure shows the ratio of the mean search time for a particle which starts its search unbound, but in the target blob, with the mean search time for the conventional random starting position. Dark regions in this figure correspond to similar search times, while bright regions show an acceleration due to the better starting position.

We observe just like in Fig. 4.5 that there are two main regimes. In a wide range of parameters the mean search time does not change much if the starting position is already close to the target. However, for increasing values of the non-specific association rate k_{ass} , there is a rather sharp transition to a regime where the local searcher finds the

⁷This will be discussed in more detail in the following chapter.

target much faster. This is best understood by looking at the relation between k_{ass} and p_r , see the discussion following Eq. (4.15).

Large values of k_{ass} imply a large probability to bind to the DNA inside a blob. For a local searcher, already starting in the target blob, this implies that it will often “choose” to bind to DNA and thus have a chance to detect the target. The combined probability to detect the target before leaving the target blob will increase and the corresponding search time will remain short. Conversely, for a global searcher, a particle starting at a random position, increased values of p_r imply that it will bind repeatedly to blobs without the target on the way. These purely non-specific blobs then act as traps, slowing down the search, before it reaches the target blob, when the increased value of p_r finally becomes advantageous for this searcher as well.

By considering vertical slices through Fig. 4.9, one can obtain a pedestrian derivation of the above mentioned independence of the search time on the starting position in 3D. The higher the dissociation rate k_{off} becomes, the more time will be spent with three-dimensional diffusion. Accordingly, for increasing values of k_{off} the search times for global and local searches approach each other.

Implications of possible anomalous transport mechanisms Within our model, the increase in the target detection rate at conditions which we deem to be biologically relevant seems to be small. This is in line with the model of “geometry-induced kinetics”, introduced by Olivier Bénichou and co-workers who showed that the starting position of a searching particle is rather unimportant, if the medium is explored in a “non-compact” fashion [193, 194].

As mentioned in the introductory section of this chapter, the motion of the lac repressor was modelled as normal Brownian motion. While in normal diffusion the mean-squared displacement of a particle grows linearly in time, $\langle x^2(t) \rangle \propto t$, subdiffusion is characterised by the behaviour $\langle x^2(t) \rangle \propto t^\alpha$ with $\alpha < 1$. Given that for several particles it was shown that they perform subdiffusion in a crowded environment [93], we shortly comment on implications of such transport mechanisms on the protein-DNA association. The sub-linear dependence on t for subdiffusion implies that the probability distribution of a particle spreads more slowly. One negative consequence of this is obvious: motion to a far-away place is slowed down. This seems to make an efficient search impossible. However, there is the other side of the coin: particles stay longer at (or close to) their initial position [195]. Thus, if a subdiffusive particle has just dissociated from the target, it does not leave the vicinal region too quickly. As discussed in the literature recently, in such a situation subdiffusion can even become beneficial for certain search processes [195, 196]. However, single particle tracking for small proteins appears to be not yet possible experimentally.

4.6 Summary and outlook

The version of the facilitated diffusion model presented in this chapter was specifically designed to describe the situation in a living bacterial cell. Therefore the simple DNA conformations studied in chapter 3 in terms of GFDM were replaced by a more realistic model. Based on the experimental observation that the genome consists of structural

subunits, DNA was modelled as a SAW of DNA blobs in the finite cell volume [191]. In each of these blobs a local facilitated diffusion model was solved in order to obtain the system's global target association rate.

As in conventional facilitated diffusion models, it was found that the search time depends non-monotonically both on the non-specific dissociation rate and the complementary association rate. This illustrates again that both the one-dimensional and the three-dimensional search phase on their own are disadvantageous, while their combination can yield an efficient search process. An investigation of the simulation results showed that the particle experiences the nucleoid as an effective medium through which it diffuses with a reduced three-dimensional diffusion coefficient.

It was found that details such as the individual conformation of DNA or parameters describing the structure did not influence the mean search time significantly. Moreover, the obtained search time was close to the experimentally measured value. Interestingly, within our model starting the search process already in the target blob did not greatly reduce the search time. This was due to the fact that for typical parameter values most of the search time was spent returning to the target blob after dissociating from it without having detected the target. Accordingly, getting rid of the initial search phase which leads to the first encounter with the target only modestly increases the target association rate. One has to note, however, that this "local" setting is rather soft. The particle was assumed to be in the three-dimensional space within the target blob which contains several kbps of DNA.

In the following chapter we focus on a different question. Concurrently, we take into account two aspects which were neglected in this chapter in order to reduce the number of parameters: the occurrence of search and recognition state of the TF and the presence of other non-specifically bound particles on the DNA.

5 Target search in a real sequence

The way how the search of a protein for a target sequence on DNA was treated in the two previous chapters was based on several assumptions concerning the interaction energy between TF and DNA. In the *in vivo* model we assumed that there is just one binding mode and that on encounter the target is detected with certainty. Conversely, in the model presented in chapter 3 it was assumed that there are two binding modes but that the switching between these two occurs independently from the underlying sequence. Both assumptions allowed a thorough theoretical treatment, but recent experimental findings gave new insights. Indeed, blind switching cannot be a good strategy, since such a change of conformation towards the recognition mode should only happen when the underlying sequence is similar to the target sequence.

In the present chapter we use an approach which is based on the one introduced in reference [197]. Therein the search of a two-state protein in a random energy landscape was described and effects of correlated rebinding were taken into account [197]. We develop this model considerably further by using the real nucleotide sequence of a common laboratory strain of *E. coli*. Furthermore to model the situation in a living cell, we take the presence of other non-specifically bound proteins into account.

Apart from the “searching” protein we usually study, in a living bacterial cell there is a plethora of other macromolecules both in the cytoplasm and bound to DNA. That molecules in the cytoplasm cause a reduced mobility of the searching particles in 3D was already taken into account in the previous chapter. However, a large number of DNA-bound proteins will obviously also influence the sliding motion of a searching TF. First, they could block the target by binding to it, and second they constitute “roadblocks” if one assumes that a sliding TF cannot by-pass them [141].

Already in 1974, James D. McGhee and Peter H. von Hippel noticed the importance of binding of large ligands to a DNA chain [198]. They calculated the probability that a gap of certain length remains free as a function of the size and number of non-specifically bound proteins [198]. These calculations of this “parking lot problem” were recently renewed by Henrik Flyvbjerg et al., leading them to the negative conclusion which we already mentioned that it is “impossible, *in vivo*, for any protein to rapidly conduct random one-dimensional diffusional sliding along DNA over a distance exceeding a few tens of base pairs” [140]. Based on the more recent experimental findings that the sliding lengths of lac repressor in a living *E. coli* cell are just a few dozen bps [104], this might not be too harsh a restriction, but it affects the search process nevertheless. In fact, Gene-Wei Li, Otto G. Berg and Johan Elf were the first who explicitly coupled these one-dimensional crowding effects to a facilitated diffusion model [141]. Most importantly they found that bound roadblocks reduce the effective antenna size from which the searcher can slide into the target [141]. But at the same they decrease the amount of non-specific DNA which traps the searcher on its way to the target [141].

In the first part of this chapter we focus on the final stage of the search for the target,

i.e. when the TF has just landed at a position near the main operator $O1$. Here the presence of other non-specifically bound proteins becomes important. The result of both a numerical model and an approximate theoretical model in this local setting will be discussed in section 5.2. In section 5.3 the full search model is introduced including looping between two operators. Finally, section 5.4 presents the results of this model.

5.1 Search in the target region

In the present model, we assume that a certain number of non-specific proteins has bound to the DNA and that they cannot be by-passed by the searching protein. For the sake of simplicity, we assume that these “blockers” are immobile, although including their mobility might be important in some cases [141,156]. Thus, when speaking of the target region we mean the possible binding positions between the nearest blocker to the left and the nearest blocker to the right of the target¹.

Importantly, unlike in the two previous chapters we study a discrete model. Within the model, at each of the discrete binding sites the particle can be either in the search or in the recognition mode. While in the loosely bound search state at position i , four events are possible: the particle can slide to the neighbouring positions $i-1$ or $i+1$ while staying in the search mode. Furthermore it can dissociate or it can perform a conformational switch to the recognition mode at position i . As a boundary condition we apply that a particle which is at the end of the target region, i.e. at a site adjacent to a blocker, and which tries to move further in that direction will be kept at the present position. In the recognition mode, we assume that binding is so tight that no sliding is possible, such that returning to the search state at the same position is the only option. All dynamic processes within this discrete state model are shown in the schematic figure 5.1.

In this simplified version of the search model, we are only interested in the probability to detect the target before dissociation from the DNA occurs. Thus, the simulation is stopped when either the particle dissociates from DNA or when it detects the target. In the present case, target detection constitutes in making a conformational switch from the search mode to the recognition mode at the position of the main operator $O1$. This deviates from the treatment of Hu et al., where an additional detection step is needed at the position of the target [167,199]. It also deviates from the treatment in terms of GFDM as presented in chapter 3, where target detection occurred via slow sliding in the recognition mode.

In simulating the system we apply the standard Gillespie algorithm [200], as the conventional model for activated transport, compare [126,201]. Thus, the rates at which the events mentioned above occur depend on the energetic barriers between the corresponding states. In general terms. the rate k_{ab} for going from state a (energy E_a) to state b (energy E_b) is given by

$$k_{ab} = \lambda_0 \cdot \begin{cases} \exp(-(E_{ba} - E_a)/(k_B T)) & \text{if } E_{ba} > E_a, \\ 1 & \text{else,} \end{cases} \quad (5.1)$$

¹Here we make the same assumption as others [122], that the target is not occupied by non-specific blockers. Of course, in reality this can happen, but in the case of immobile blockers, there is no point in simulating a search process in which the target cannot be found.

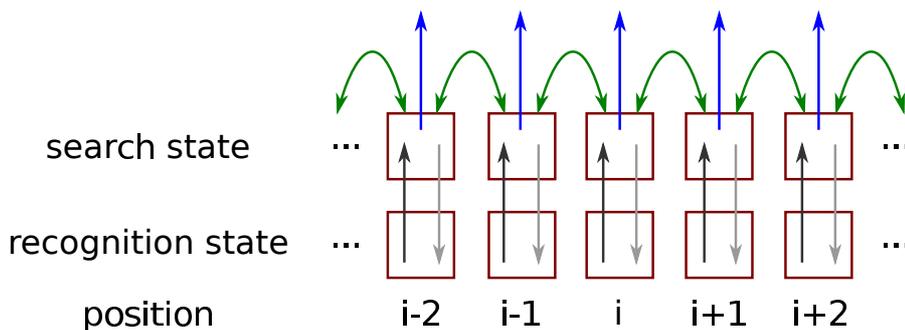


Figure 5.1: Scheme of the search process within the discrete state model. At any position the TF can be in the search state (upper row) or in the recognition state (lower row). Interconversion between these states occurs at rate k_{rs} (from recognition to search state - dark grey arrows) and at rate k_{sr} (from search to recognition state - light grey arrows). Note that these rates depend on the underlying sequence and can be different for any site. Only in the search state, the TF can slide (with rate Γ) to neighbouring sites as depicted by the green arrows. From the search state dissociation to the bulk solution (with rate k_{off} - blue arrows) is possible.

where E_{ba} denotes the energetic barrier between these two states. The inverse of the attempt rate, λ_0 , denotes the fundamental time scale of the system.

In the following we assume that the binding energy in the search state, E_s , is the same for all binding positions. This assumption which can be traced back to the classical work of Winter et al. was rationalised via Monte Carlo simulations by Dahirel et al. [59, 89]. Besides, a similar energy landscape was discussed in [122].

E_s and all other energies are measured relative to the energy of the unbound state which is set to zero. We assume that the energetic barrier, E_{bs} , to be crossed for sliding to a neighbouring position is sequence-independent as well. This is in contrast to the treatment of Zabet and Adryan, in which the specific binding energy has to be brought up for each sliding step [155].

However, we assume that the binding energy in the recognition mode and the energetic barrier which has to be crossed to get there from the search state at the same position depend on the underlying nucleotide sequence. The following treatment is based on the assumption that this switching should occur more easily when the sequence is similar to the target sequence. We will now detail how this “similarity” can be quantified.

5.1.1 Score matrix

In Fig. 1.4 we already saw qualitatively that the three naturally occurring operators are similar and have a length of 21 base pairs². The DNA can be considered as a one-dimensional array, in which $w = 21$ adjacent nucleotides correspond to a possible binding site for the lac repressor. Based on the knowledge that the repressor binds strongly to

²Note that the symmetric operator O_{sym} is one nucleotide shorter and cannot be dealt with using the methods described in this section.

its three operators, one usually calculates the position weight matrix (PWM) score, S for each of these putative sites. The higher the score is, the stronger the repressor will bind to the corresponding site. It is calculated via [202]:

$$S = \sum_{j=1}^w \log_a \left(\frac{1}{p(l_j)} \frac{f_{l_j,j} + s(l_j)}{N_{\text{bs}} + \sum_b s(b)} \right). \quad (5.2)$$

In this general form p_b denotes the background frequency of base b , l_j denotes which nucleotide is found on position j of the putative sequence and N_{bs} the number of binding sites. Finally, $s(b)$ is a pseudo-count function in order to avoid a divergence of the term when $f_{l_j,j} = 0$. In the present case, we have $N_{\text{bs}} = 3$ and we apply the simple convention used by José M. G. Vilar in order to obtain comparable results [203]. Namely, we use the natural logarithm, i.e. $a = e$, for the sake of simplicity a homogeneous background sequence, $p(l_j) = 0.25 \forall l_j$ and $s(b) = 1 \forall b$. Then Eq. (5.2) simplifies to

$$S = \sum_{j=1}^{21} \ln (4(f_{l_j,j} + 1)/7). \quad (5.3)$$

In agreement with the result of Vilar we obtain the following scores for the three known natural operators [203]: $S_{O3} = 10.95$, $S_{O2} = 12.17$ and $S_{O1} = 13.38$. But how can this score be connected to the binding energy at the corresponding position?

5.1.2 Relation between scores and energies

In order to establish such a connection, in the following all energies and scores are related to a hypothetical average binding position which has the mean score of all binding positions, $\langle S \rangle$, and is characterised by the specific binding energy, E_r . Introducing $\Delta S_i = S_i - \langle S \rangle$, the difference between the score at position i and the average score and assuming a linear relationship between scores and specific binding energies (compare [62, 204]), one obtains:

$$E_{r,i} = E_r + \gamma \times \Delta S_i. \quad (5.4)$$

Here, γ is simply a proportionality factor which can be deduced from experimental data. Based on the assumed linear relation between scores and energies, we obtain the correct order in terms of binding strength as measured in the 1990s by Stefan Oehler et al. or more recently by Hernan G. Garcia et al. [205, 206].

As a central element of this model, we furthermore assume that the energetic barrier between the energy in search and recognition mode at position i is given by a similar relation:

$$E_{\text{bc},i} = E_{\text{bc}} + \alpha \times \gamma \times \Delta S_i. \quad (5.5)$$

In analogy to Eq. (5.4), E_{bc} denotes the mean value of the barrier height between search and recognition state. The parameter α will be called “volatility parameter” in the following. This can be understood by considering two specific cases:

- $\alpha = 0$ implies that the barrier height is independent from the nucleotide sequence. In particular, at the target site it is as high as at any other binding site. Thus, in this limiting case the particle probes blindly for the target.

- $\alpha = 1$ describes the situation in which the barrier height varies as much around its mean value as the specific binding energy. Accordingly, the barrier height at the target site is the smallest one in the system and an *induced switch* mechanism is at work.

In the following we use intermediate values of α to interpolate between these two limiting cases. The reasons why we call α volatility parameter is that the higher the value of α is, the more probable it becomes that the TF probes for the target at a non-target site. Note that in principle we could also study values of α larger than unity or smaller than zero. We do not apply them, however, since they appear to be unphysical. Recently, it was found experimentally that the association rates to O_{sym} and $O1$ are similar, while dissociation occurs more rapidly from the natural operator $O1$ [106]. In terms of our present model this implies rather large values of α , since then the barrier at the target site is small for both operators and the dissociation rate depends on the absolute value of the binding energy which is minimised for O_{sym} .

In addition, when the barrier height at a position calculated via Eq. (5.5) is smaller than the energy of the final state, the latter is inserted into Eq. (5.1) for calculating the corresponding rate in the Gillespie scheme.

5.1.3 Reference set of parameters

In the following we state the reference set of parameters that we employ. Most importantly, we take an excerpt of the real nucleotide sequence of the *E. coli* strain K-12 MG 1655: the base pairs 359,990-370,010 which comprise the genes *lacA*, *lacY* and *lacZ* and the three natural operators of the *lac* repressor were imported from ecocyc.org, i.e. Kessler et al. [30]. As the binding motif of the *lac* repressor has a length of 21 nucleotides, this implies that our system comprises $10,021 - 21 + 1 = 10,001$ possible binding positions in both orientations. A histogram of scores obtained for this set of putative sites is shown in Fig. 5.2.

While most of the scores in both orientations have similar intermediate values, the three scores corresponding to the three natural operators can be found at the upper end of the score distribution. A thorough discussion of the influence of such gaps on the stability of binding can be found in Sheinman et al. [124].

Here and throughout this chapter, energies are measured in units of $k_B T$. We chose $E_r = 0$, $E_s = -7$, $E_{\text{bs}} = -6$ and $\gamma = -1.3378$ such that the energy difference between the binding energy at the main target site and the one in the search state is 15.3 [206]. Besides, we put $N_{\text{block}} = 71$ non-specific binders as roadblocks at non-overlapping, but otherwise random positions. This yields the same level of site occupation as in Tabaka et al. [142]. Finally, each time 50,000 simulation runs are performed for each blocker configuration and in each case it is recorded whether and when the target is reached. This result is then compared to a theoretical model which we describe in the following subsection.

5.1.4 Theoretical model

To determine the events in the target region, we focus on the N binding positions around the target which lie between the two blockers which are nearest to the target. We only

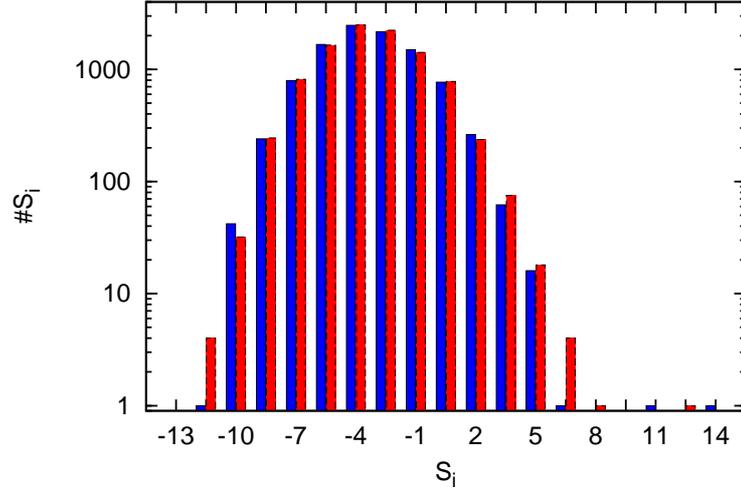


Figure 5.2: Histogram of score values calculated with Eq. (5.3) for the 10,001 putative binding positions surrounding the main operator in E.coli K-12 MG 1655. Red and blue bars refer to the two different orientations of binding positions. At the upper end of the spectrum the three natural operators can be found. Note the logarithmic scaling of the y -axis.

consider the binding energies in the correct orientation and N denotes an odd number, such that there is a central site, but the conclusions we draw do not depend on this assumption. The situation is simplified with respect to the model which we simulate numerically, in that we assume that the rates for conformational switching events are equal at all positions but the target. This simplification enables an analytical solution of the problem and in the following we will assess if this still accurately describes the situation.

We study the probability densities of particles on base pair j at time t both in the search state, denoted by $c_{N,j}(t)$, and in the recognition state, $p_{N,j}(t)$. Assuming that the target is on node m , they are subject to the following differential equations:

$$\begin{aligned} \frac{\partial c_{N,j}}{\partial t} = & \Gamma [c_{N,j-1}(t) + c_{N,j+1}(t) - (2 - \delta_{j,1} - \delta_{j,N})c_{N,j}(t)] - k_{\text{off}}c_{N,j}(t) \\ & - [k_{\text{sr}} + \delta_{j,m}(k_{\text{st}} - k_{\text{sr}})] c_{N,j}(t) + k_{\text{rs}}(1 - \delta_{j,m})p_{N,j}(t), \end{aligned} \quad (5.6)$$

where Γ denotes the sliding rate to the next site (the discrete analogue to D_1), k_{off} the dissociation rate and k_{rs} the switching rate from the recognition state to the search state (at all non-target sites). Finally, k_{sr} denotes the rate of changing from search state to recognition state away from the target and k_{st} the one for doing this at the target site. Obviously, only densities at neighbouring sites are coupled, such that the matrix to be solved is tridiagonal.

The corresponding equations for the density probabilities in the search state follow as:

$$\frac{\partial p_{N,j}}{\partial t} = [k_{\text{sr}} + \delta_{j,m}(k_{\text{st}} - k_{\text{sr}})] c_{N,j}(t) - k_{\text{rs}}(1 - \delta_{j,m})p_{N,j}(t). \quad (5.7)$$

If the particle's initial position is in the search state at an arbitrary position in the interval, initially there is no probability density in the recognition mode. Then, it is advantageous to go to Laplace space with respect to time, which yields

$$\tilde{p}_{N,j} = \tilde{c}_{N,j} \frac{k_{\text{sr}} + \delta_{j,m}(k_{\text{st}} - k_{\text{sr}})}{u + (1 - \delta_{j,m})k_{\text{rs}}}, \quad (5.8)$$

where here and in the following the explicit dependence on the Laplace variable u was omitted. In other words, at all non-target sites, one has $\tilde{p}_{N,j \neq m} = k_{\text{sr}} \tilde{c}_{N,j \neq m} / (u + k_{\text{rs}})$, while at the main operator, $\tilde{p}_{N,m} = k_{\text{st}} \tilde{c}_{N,m} / u$, is valid.

The Laplace transform of the flux of particles to the target follows readily:

$$\tilde{j}_{N,m} = k_{\text{st}} \tilde{c}_{N,m} = u \tilde{p}_{N,m}. \quad (5.9)$$

We now focus on the situation when the particle starts its search on site n which is indicated by the third index of the quantities below. Later on, we will assume that the search in the target region is started on a random position and thus we will average over all starting positions. Then we will drop this third index again. It is convenient to introduce the auxiliary function, $\zeta(u) = u + k_{\text{off}} + k_{\text{st}}$. For the flux into the target on site m in a system of N sites, where the particle starts on node n , we obtain

$$\tilde{j}_{N,m,n} = k_{\text{st}} \tilde{c}_{N,m,n} = u \tilde{p}_{N,m,n} = \frac{\sum_{i=0}^{N-1} \hat{k}_{\text{st}} a_{i,N,m,n} \hat{\Gamma}^i}{\sum_{i=0}^{N-1} \left((\hat{\zeta} - 1) a_{i,N,m,m} + \sum_{n'=1}^N a_{i,N,m,n'} \right) \hat{\Gamma}^i}. \quad (5.10)$$

Here the hat over a function means that the corresponding quantity was divided by yet another auxiliary function, defined as $\xi(u) = k_{\text{off}} + u[1 + k_{\text{sr}}/(u + k_{\text{rs}})]$. As seen throughout this work, for our present purposes the Laplace transforms contain the most important information at the origin, $u = 0$. In particular, one has $\xi(u = 0) = k_{\text{off}}$. Therefore, as another short notation, a bar over a quantity denotes that it was divided by k_{off} . Namely, we will use $\bar{\Gamma} = \Gamma/k_{\text{off}}$ and $\bar{k}_{\text{st}} = k_{\text{st}}/k_{\text{off}}$. The expansion coefficients in Eq. (5.10) are defined as

$$a_{i,N,m,n} = \begin{cases} \sum_{j=\max\{0, n+i-N\}}^{n+\min\{-1, i-m\}} \binom{2(n-1)-j}{j} \binom{2(N-m)-(n-m+i-j)}{n-m+i-j} & \text{if } n \leq m, \\ \sum_{j=\max\{0, i-n+1\}}^{\min\{N, i+m\}-n} \binom{2(N-n)-j}{j} \binom{2(m-1)-(m-n+i-j)}{m-n+i-j} & \text{if } m \leq n. \end{cases} \quad (5.11)$$

In the following we consider the situation of a random starting position and accordingly we drop the third index again:

$$\tilde{j}_{N,m} = \frac{1}{N} \sum_{n=1}^N \tilde{j}_{N,m,n}. \quad (5.12)$$

As in the previous chapters, the 0th order of \tilde{j} yields the probability to detect the target before dissociation occurs, $p_t(N, m)$, where the first argument denotes the number of sites in the target region and the second one on which one of those the target resides. Rearranging the terms, one obtains:

$$[Np_t(N, m)]^{-1} = \bar{k}_{st}^{-1} + [1 + G(\bar{\Gamma})]^{-1}. \quad (5.13)$$

Here the auxiliary function G , which is independent of k_{st} and of $k_{rs/sr}$ is given by

$$G(\bar{\Gamma}) = \frac{\sum_{i=0}^{N-1} \left(\sum_{n \neq m} a_{N,i,m,n} \right) \bar{\Gamma}^i}{\sum_{i=0}^{N-1} a_{N,i,m} \bar{\Gamma}^i}. \quad (5.14)$$

Thus, G solely depends on the renormalised sliding rate, $\bar{\Gamma}$, the size of the system and the position of the target. Eq. (5.13) together with Eq. (5.14) represents the central result of this chapter.

Writing $\varepsilon = 1/(2\bar{\Gamma})$ and $y = 1 + \varepsilon - \sqrt{\varepsilon(2 + \varepsilon)}$, the second term on the right hand side of Eq. (5.13) is equivalent to the result obtained by Kolomeisky and co-workers [88,207,208]:

$$(1 + G(\bar{\Gamma}))^{-1} = \tanh\left(\frac{\ln[y]}{2}\right) \frac{\cosh([N - (m - \frac{1}{2})] \ln[y]) \cosh([m - \frac{1}{2}] \ln[y])}{\cosh(N \ln[y]/2) \sinh(N \ln[y]/2)}, \quad (5.15)$$

or upon substituting $\varsigma = -\ln[y]$ [209]:

$$(1 + G(\bar{\Gamma}))^{-1} = \tanh\left(\frac{\varsigma}{2}\right) \frac{\cosh([N - (m - \frac{1}{2})]\varsigma) \cosh([m - \frac{1}{2}]\varsigma)}{\cosh(N\varsigma/2) \sinh(N\varsigma/2)}. \quad (5.16)$$

In other words, our Eq. (5.13) generalises the result of Kolomeisky and co-workers to the biologically relevant regime when not every target encounter leads to detection. In the specific case of a target on the central node, i.e. $m = (N + 1)/2$, and when $\bar{k}_{st} \gg 1$, one obtains

$$p_t(N, (N + 1)/2) = \frac{\tanh\left(\frac{N}{2}\varsigma\right)}{N \tanh\left(\frac{1}{2}\varsigma\right)}, \quad (5.17)$$

which has the typical tanh-dependence of sliding models, compare Eqs. (2.30) and (4.4) or reference [128].

Introducing q , which denotes the ratio of the conformational switching rates, $q = k_{sr}/k_{rs}$, one obtains for the conditional mean search time:

$$k_{off} \tau_{N,m} = \frac{1 + (1 + q) \left[G + \frac{\bar{k}_{st} \bar{\Gamma}}{1 + G} \frac{\partial G}{\partial \bar{\Gamma}} \right]}{\bar{k}_{st} + 1 + G}. \quad (5.18)$$

For a target on the middle site and considering the limiting case of vanishing dissociation rate, the function G approaches $N - 1$ and its derivative with respect to $\bar{\Gamma}$ approaches $N(N^2 - 1)/(12\bar{\Gamma}^2)$. This leads to

$$\tau_{N, \frac{N+1}{2}} = \frac{1}{k_{st}} + (1 + q)(N - 1) \left[\frac{N + 1}{12\bar{\Gamma}} + \frac{1}{k_{st}} \right]. \quad (5.19)$$

In this equation, the role of the ratio q is most apparent. The fact that the particle can switch to the recognition mode at sites away from the target implies that the second term, which roughly describes the time spent on the way to the target, has to be multiplied by $1 + q$. Stated differently, this simply means that probing for a target where it cannot be found, slows down the search in comparison to an “informed” particle which only probes at the target site. However, the target detection probability as given by Eq. (5.13) does *not* depend on these switching rates.

When q approaches zero, that is when there is no switching to the recognition mode on non-target sites, our result converges to results obtained in the theory of incoherent exciton hopping, where the impurity quenching of molecular excitons was studied [210, 211], yielding:

$$\tau_{N, \frac{N+1}{2}}^{q=0} = \frac{N}{k_{\text{st}}} + \frac{N^2 - 1}{12\Gamma}. \quad (5.20)$$

In terms of exciton transport τ represented the mean de-excitation time of the molecular aggregate [210]. In the following numerical results for the search in the target region obtained within the Gillespie scheme are compared to this simplified theoretical model.

5.2 Results for the search in the target region

5.2.1 Probability to detect the target

For 500 configurations it was simulated whether the target was detected on DNA or the particle dissociated first. The resulting probability to detect the target prior to dissociation is shown in Fig. 5.3 for three different values of α and as a function of the size of the target region, N . This size N is simply the number of possible binding sites.

Blue symbols correspond to simulations with $\alpha = 0.1$, black ones to $\alpha = 0.3$ and red ones to $\alpha = 0.5$. The lines of corresponding colour show results obtained with the simplified theoretical model as given by Eq. (5.13), where full lines are for a target exactly in the centre of the region and dashed lines for the opposite case when the target is at the boundary of the region.

For the two larger values of α the results of the simulation are in complete agreement with the simplified theoretical model. Also in the case of $\alpha = 0.1$ (blue line and symbols) the simulation results are close to the theoretical result. The small deviations are simply due to the finite sample size. Overall, for all values of α the target detection probability decreases with increasing values of N . This is expected as in a more spacious system with homogeneous initial distribution of particles the mean distance to the target is larger than in a smaller system. And a larger distance to the target implies more chances or risks to dissociate.

At the same time, at a fixed value of N larger values of α guarantee a higher detection probability. Clearly, at small values of α the barrier at the target is only marginally smaller than at any other site, resulting in many undesired overshoots at the target. However, as implied by the form of Eq. (5.1) increasing α above a certain threshold does not increase the detection probability. If there is no longer a barrier to be crossed, target detection on encounter occurs with the attempt rate λ_0 no matter if α is increased even further.

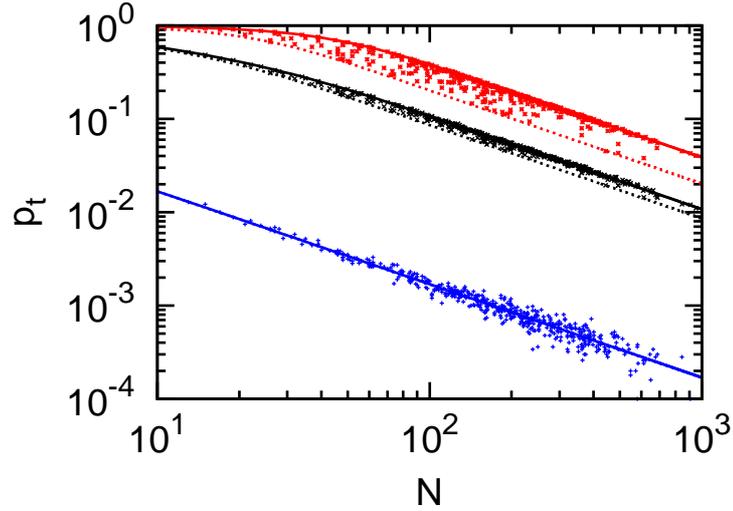


Figure 5.3: Probability to detect the target prior to dissociation as a function of the target region size N . Results from numerical simulations (symbols) are compared to the theoretical result, Eq. (5.13). Parameters: $E_s = -7$, $E_{bs} = -6$, $E_r = 0$ and $E_{bc} = 4$. Dashed lines correspond to a target at the boundary of the system, while full lines to a centred target. Colours: blue ($\alpha = 0.1$), black ($\alpha = 0.3$) and red ($\alpha = 0.5$).

A more subtle influence of different values of α is seen when comparing results for different target positions, but otherwise fixed parameters. To do this dashed and full lines of matching colour in Fig. 5.3 have to be compared. The fact that for the two larger values of α , a centred target leads to a more reliable detection is again explained by the reduced mean distance to the target in that case. For $\alpha = 0.1$ (blue lines and symbols) no such difference is observed. Here, the detection probability on an individual encounter is so small that the optimised position of the target does not ameliorate the search.

Thus, in terms of the target detection probability the simplified theoretical result nicely agrees with the numerical simulations. However, as we will see below in the case of the detection *time* the assumption that all non-target sites can be replaced by an average site will no longer be valid.

Influence of the target region size In the previous subsection it was found that the target detection probability p_t decreases with N . However, it also makes sense to consider the product $N \cdot p_t$. This is roughly the number of nucleotides from which target detection via sliding is certain and is similar to the antenna length encountered in previous chapters. While in general a larger target region will attract more particles, the individual detection probability is smaller.

In analogy to the discussion following Eq. (4.4), the numerator in Eq. (5.17) converges to 1 for large N . Thus, Np_t converges to the constant value $\coth(\varsigma/2)$. This implies that a target region which is much larger than the sliding length will not lead to an increased

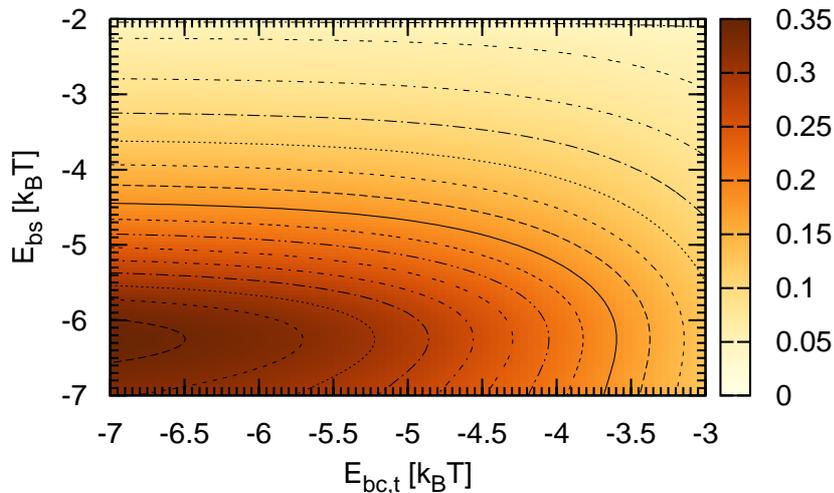


Figure 5.4: Target detection probability as a function of the energetic barrier for sliding, E_{bs} , and the energetic barrier for target detection, $E_{bc,t}$. Parameters: $N = 101$, $m = 51$, $k_{\text{off}} = \exp(-7)$.

antenna size, since particles are more likely to dissociate before they arrive at the target.

5.2.2 Impedance matching

It was noted by Hu and co-workers and similarly by Bénichou et al. that for a finite target detection rate too high sliding rates become disadvantageous [167,212]. While for an infinite detection rate, it is obvious that having a high mobility in 1D implies that the target will be found quickly and reliably, in the finite case fast sliding can lead to overshooting the target. We invoked this effect which is similar to impedance matching already to explain the features of Fig. 5.3, but present it in more detail in Fig. 5.4. Based on the observation that the theoretical result given by Eq. (5.13) nicely reproduces the numerical results, we plot the target detection probability for various values of the energetic barrier which has to be crossed for sliding and for target detection. We study a system with $N = 101$, a centred target ($m = 51$) and $k_{\text{off}} = \exp(-7)$.

As is apparent from Fig. 5.4, in a horizontal slice the detection probability always decreases from left to right. Thus, it is always beneficial for a reliable target detection to have a low barrier to the recognition mode at the target site. However, the detection probability depends non-monotonically on the sliding rate. For any fixed value of the barrier for target detection, there is an optimal barrier for sliding. The higher the rate at which particles are drawn into the target, the higher the optimal sliding rate is. Roughly, one can say that it is advantageous to supply the target with exactly the amount of particles it can absorb, but not more.

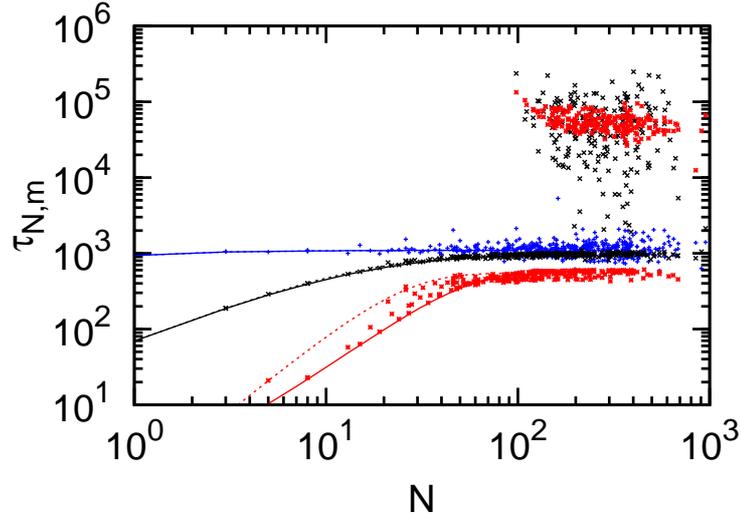


Figure 5.5: The conditional target detection time as a function of the length of the target region. Lines: theoretical model (full: centred target, dashed: target at the boundary). Symbols: numerical results. Colours as in Fig. 5.3. Note that due to the presence of the auxiliary operator $O3$ in the target region a second branch of results appears for large values of N .

5.2.3 Conditional target detection time

We now turn our focus to the mean first passage time to the target. Unlike in the previous chapters, this is now a conditional detection time. The corresponding results are presented in Fig. 5.5, where symbols are results obtained in numerical simulations and lines show the theoretical results. The meaning of colours and line types is the same as in Fig. 5.3, while the range of N has been extended for a better visibility of the trends.

For $N \lesssim 100$, the simplified theoretical model again is able to reproduce the numerical results reliably. Particles characterised by a larger value of α on average detect the target faster. Even more striking than for the detection probability, for the conditional detection time only in the case $\alpha = 0.5$ (red) the position of the target makes a difference. This can be seen by observing that in both other cases the straight lines and the dashed lines lie on top of each other. For $\alpha = 0.5$ a centred target leads to a faster target detection, again due to shorter mean distances to the target.

Conversely, for $N \gtrsim 100$ we observe that there are two branches of conditional search times. While the lower one can be considered to be the continuation of the results for short target regions, the upper branch shows target detection times which are considerably larger. These results are due to the auxiliary operator $O3$ which has the same orientation as $O1$ and is at a distance of 92 nucleotides from the main operator. Thus, whenever the target region extends over more than 92 nucleotides there is the possibility that both $O1$ and $O3$ are in the target region. If the TF finds and “detects” the auxiliary operator, this represents a trap on the way to detecting the main target. Since the auxiliary operator conveys a rather strong binding site, this severely impacts the detection

time leading to the upper branch in Fig. 5.5. Obviously, for such a strong binding site it is no longer appropriate to assume that it can be represented by an average non-specific binding site.

The *probability* to detect the target is, however, not influenced by the presence of the auxiliary operator (compare Fig. 5.3). As a side remark, we note that since we assume that during sliding the TF cannot change its orientation in this simplified model the presence or absence of *O2* in the target region is unimportant.

When studying the features of the upper branch of detection times in Fig. 5.5, two facts are striking. First, there are hardly any detection times for $\alpha = 0.1$. In Fig. 5.3 it can be seen that already the probability to detect the main operator is rather small for this value of α . This is even more pronounced in the case of the weaker operator *O3*. Thus, for these particles it is very unlikely that they will be trapped at *O3*. However, this comes at the price that the detection of the main operator will be rather unlikely as well. In fact, in subsection 5.2.4 we will study the ability of the TF to distinguish between the two operators.

The second striking fact, is that for $\alpha = 0.5$ and intermediately sized target regions ($100 \lesssim N \lesssim 150$) the target detection time slightly decreases for increasing values of N . Again this is due to differences in the mean distance from the target. In regions which are only slightly longer than 92 nucleotides and which comprise both *O1* and *O3*, these two operators both lie near to the two boundaries, which was already diagnosed as leading to longer search times. In longer target regions more central operator positions are possible, leading to reduced search times. For even longer regions, the search time levels off but showing large scatter.

5.2.4 Probability of first detecting *O1*

The volatility parameter α can be considered to be a measure to make sure that a switch of conformation only occurs when the underlying nucleotide sequence is equal or similar to the target sequence. In the following we study whether it is possible to choose α such that it is more probable to detect the main operator *O1* than the auxiliary operator *O3*. We define the preference of *O1* over *O3* as the probability of detecting *O1* first minus one half. Thus, the preference will attain values between minus one half and plus one half and when it is more probable to detect *O1* first, this preference will be positive.

Apart from the binding strength, the geometry of the target region also plays a role. For example, in the previous subsections it was shown that a central position within the target region is beneficial. To compare the centrality of both operators we introduce the variable x defined as:

$$x = |x_3 - 0.5| - |x_1 - 0.5|. \quad (5.21)$$

Here x_i stands for the relative position of the respective operator in the target region ($i = 1, 3$). $x_i = 0$ signifies that operator i is at the left end of the target region, and $x_i = 1$ that it is at the right end. $x_i = 0.5$ refers to a central position of the corresponding operator.

We comment on three special cases for the variable x : $x = 0$ refers to the situation when both operators are at the same distance from the middle of the target region. $x = -0.5$ is obtained when *O3* lies in the middle and *O1* at the very boundary of the system. Conversely, $x = 0.5$ implies that *O1* is in the optimal central position, while *O3*

lies at the boundary. Therefore as in the definition of the preference, positive values of x imply that the position of $O1$ is advantageous.

Using the data obtained in the numerical simulations, Fig. 5.6 shows the preference as a function of x , where due to the scatter in the data we actually present a moving average of each time 21 data points. The top panel shows the data for $\alpha = 0.5$ and α decreases in steps of 0.1 each panel downwards.

Starting the discussion with the lowest panel ($\alpha = 0.1$), we observe that in this case the preference hardly depends on x , i.e. the position of both operators. However, for all values of x the main operator is preferentially detected. Since this corresponds to nearly blind searching, the energetic barrier at the auxiliary operator is nearly as large as at an average site and only at the main operator its size is slightly reduced. This preference for $O1$ in the absence of large geometrical effects is even more pronounced for $\alpha = 0.2$ and 0.3 as manifested in the larger average values of the preference.

If α is increased further, this preference is less strong, but the positioning of both operators becomes increasingly important. This can be seen from the positive slope of the curves in the upper two panels. Note that in the top panel the y-axis has been shifted and that the preference attains negative values. Thus, in this parameter range the individual detection probability on encounter is so high for both operators that it becomes more probably to detect the weaker operator first if it is on average closer to the starting position of the searcher.

Therefore, intermediate values of the volatility parameter enable the particle to discriminate between the two operators. While checking too rarely implies that one often misses the main target, checking too often leads to the detection of the auxiliary operator. However, so far we treated $O3$ as a pseudo-operator whose detection slows down the target search for the main operator. To see how $O3$ turns from a pseudo- to an auxiliary operator [44], we have to take events after the dissociation of a TF and its ability to form loops into account.

5.3 Full search model

In the following full search model we consider that the lac repressor in its natural tetrameric form is able to form loops between two operators. However, in our setting of immobile blockers it might be that they are occupied and cannot be bound by the repressor. Thus, we distinguish between six types of configurations. They differ mainly in whether or not $O2$ is occupied and whether or not $O3$ is free. Finally, for those configurations in which $O3$ is open for binding, it makes a difference whether it lies in the target region.

So far, we focused on the events in the target region and simulations ended whenever the target was found or the searcher dissociated. Obviously, the full model has to encompass events after dissociation from DNA. We assume that the time spent in 3D is exponentially distributed with mean time τ_b . For the distance between dissociation and re-association points we assume the following cumulative distribution:

$$C(x_{\text{jump}}) = \frac{1 - \left(\frac{x_{\text{jump}}}{x_{\text{min}}}\right)^{1-\beta}}{1 - \left(\frac{x_{\text{max}}}{x_{\text{min}}}\right)^{1-\beta}}. \quad (5.22)$$

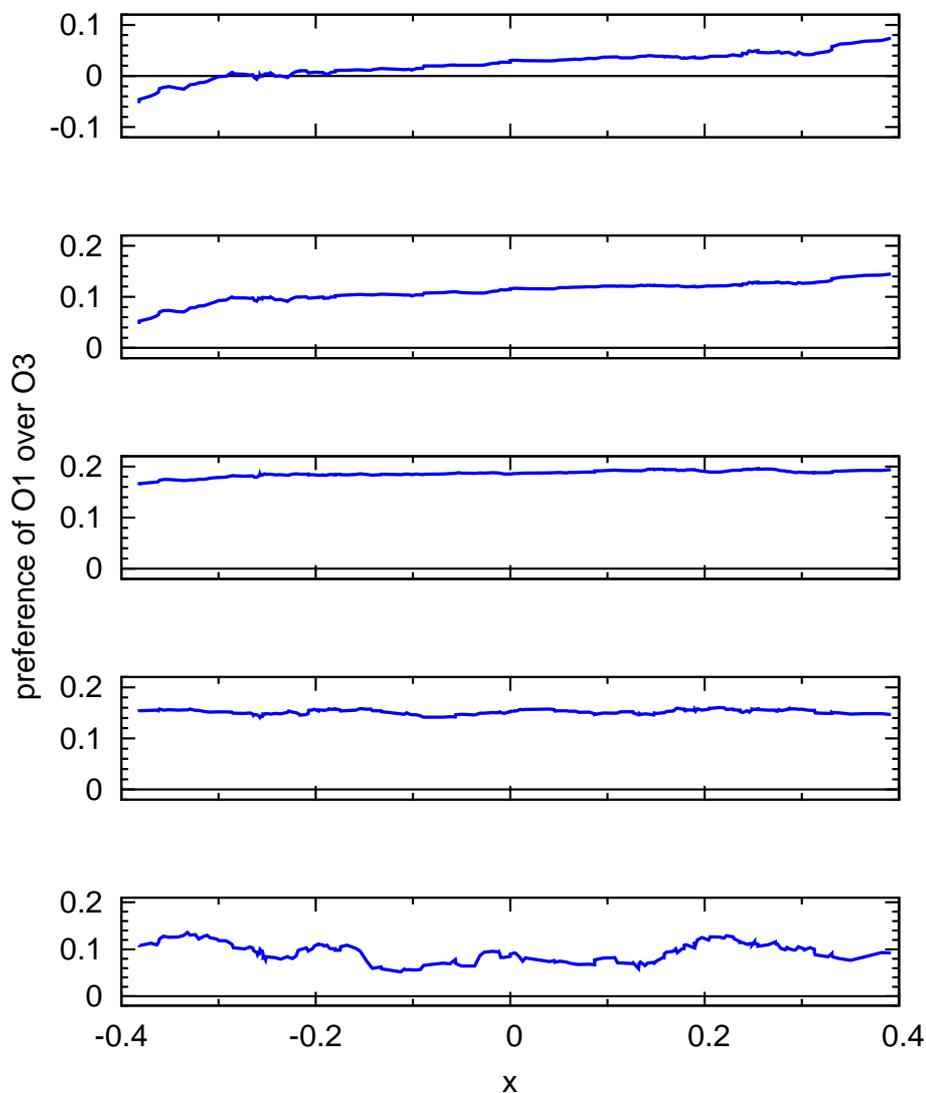


Figure 5.6: Preference of detecting the main operator $O1$ before the auxiliary operator $O3$ as a function of the proximity of both operators to the middle of the target region. Plots are moving averages for $\alpha = 0.5$ (top panel), $\alpha = 0.4$, 0.3, 0.2 and 0.1 (bottom panel). Note that the y -axis shows a different range in the top panel.

Here, x_{\min} denotes the minimum of the jump length which we fix arbitrarily to a value of 0.01. Here and throughout this chapter distances are measured in units of base pairs. Conversely, the maximum of the jump length is given by $x_{\max} = 2.3 \times 10^6$ (half the size of the *E. coli* genome) and we chose $\beta = 1.2$ as in Priest et al. [213]. Therein it referred to the power law for looping distances, but we assume here that it similarly holds for hopping and jumping lengths. If the distance drawn from this distribution implies that the particle left the range of the 10,001 binding positions we focus on, we reintroduce it at a random position in the system.

Since we are ultimately only interested in the events in the target region, we coarse-grain all events in other regions. Therefore, we simulate the mean dissociation times from all the regions which do not contain the target. The number of runs was at least twenty times the length of a region in order to obtain reliable statistics. In the main simulation, which is executed for a total time of τ_{\max} , it is then assumed that each jump ending in one of those regions consumes an amount of time which is exponentially distributed with this mean dissociation time.

Whenever the particle “detects” one of the two auxiliary operators we include the option to form a loop with the main operator (and vice versa³). In the case of looping an initiation time is drawn (exponentially distributed with mean τ_{init}). This represents the time it takes to bind non-specifically to the target region and we assume that it is equal for both auxiliary operators. Furthermore, to this initiation time we add a time lag which represents the search in the corresponding target region. This is again determined in advance by simulating the target search in the corresponding region.

Finally, the un-looping times are determined by yet another simulation in which the particle starts on the target site and the mean dissociation time from the target region is determined. Note that in this case, the TF can return to the target site several times before it eventually dissociates.

5.4 Results for the full search model

In the following we present results for one representative of each of the six classes of blocker configurations mentioned above. To ensure a comparability of the results, they were chosen to have a similar length of $N \approx 180$. Keeping the biological role of the lac repressor in mind as detailed in the introductory chapters, we now focus on the distribution of times during which the main operator is not bound by the repressor. Exactly in these time spans RNAP is able to bind to the promoter in order to express the genes. In the following we study the dependence of this quantity on the various system parameters.

5.4.1 Dependence on α

In the first case we focus on a configuration in which none of the auxiliary operators is open for binding. In this system without looping we study the dependence on the volatility parameter α . The distribution of time spans in which the main operator is

³For the sake of simplicity, we do not include direct looping between the two auxiliary operators. As we will see below, when *O3* lies within the target region, they appear nevertheless.

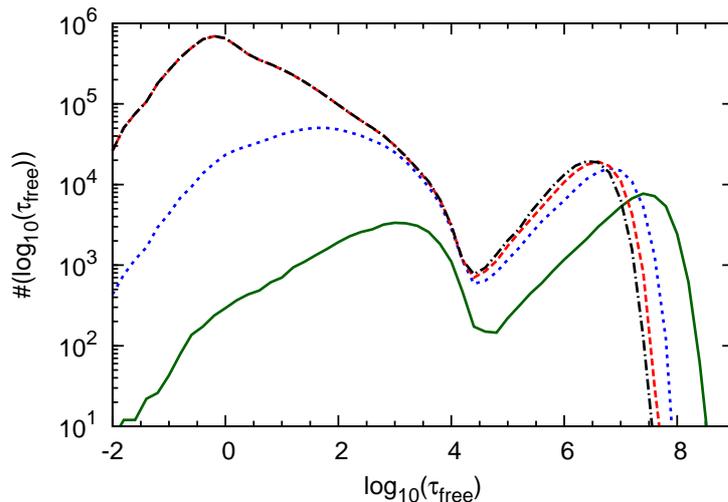


Figure 5.7: Distribution of periods in which the main operator is not bound by the repressor. Note that looping is excluded in this system and we employ: the mean time spent in 3D, $\tau_b = 50 e$, the loop initiation time $\tau_{\text{init}} = 10 \tau_b$ and the total simulation time $\tau_{\text{max}} = 3 \cdot 10^{13}$. Full dark green line ($\alpha = 0.3$), blue dashed line ($\alpha = 0.4$), red dashed line ($\alpha = 0.5$), black dot-dashed line ($\alpha = 0.6$).

not bound by the repressor is shown in Fig. 5.7. For all four values of α studied herein, two peaks are observed. When α attains larger values, the size of the first peak grows relative to the size of the second one, compare e.g. the dark green line to the black line. At the same time, the total number of events grows with α . This is simply due to the fact that τ_{max} was kept fixed in all four cases. Both peaks are shifted to shorter search times for increasing values of α , in particular in the case of the first peak.

The behaviour of the two peaks as a function of α suggests the following interpretation. The first peak describes events in which the repressor slides only shortly or not at all prior to returning to the target. Importantly, during this excursion it does not dissociate from the DNA. The second peak represents longer-lasting events in which the TF dissociates at least once.

Now, for larger values of α the individual detection probability on encounter increases. Thus, it becomes increasingly probable for the particle to quickly return to the target and if it happens, this occurs faster (compare Figs. 5.3 and 5.5). This explains the relative and absolute growth of the first peak which is also shifted to the left.

While the quick rebinding events are a result of our study, one has to keep in mind that they can be so fast that a real cell as a whole does not “perceive” them. Only if the operator remains free long enough for an RNAP molecule to bind to the promoter, the biological state of the cell changes. Stated differently, if the repressor remains associated with the target for a given time span or if it shortly dissociates away from the target, but returns before an RNAP can bind to the promoter makes no difference for the cell. Such effects were discussed in much detail in the literature, for example by Pieter Rein

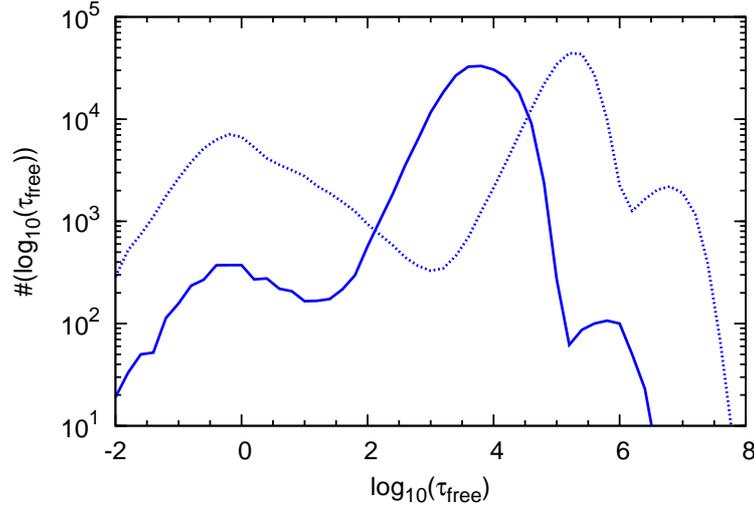


Figure 5.8: Distribution of time spans during which the main operator is not bound by the repressor in a system where the main operator and the auxiliary operator $O2$ are accessible. Thus unlike in the situation depicted in Fig. 5.7 looping between $O1$ and $O2$ is possible. Full line: $\tau_b = 50 e$. Dashed line: $\tau_b = 5000 e$. In both cases $\tau_{\text{init}} = 10 \tau_b$.

ten Wolde and co-workers [214].

5.4.2 Looping effects

Next we focus on a configuration in which looping is possible since $O2$ is accessible. We study again the distribution of times during which the main operator is free for a fixed value of $\alpha = 0.6$ and two different binding times τ_b . The results are shown in Fig. 5.8.

We first study the features of the full line which was taken with the same microscopic parameters as Fig. 5.7, compare the black line therein which also shows the case $\alpha = 0.6$. Again there are two peaks at $\log_{10}(\tau_{\text{free}}) \approx 0$ and at $\log_{10}(\tau_{\text{free}}) \approx 6$ which represent the same events as before. But between these two, there is now a third peak. It corresponds to events in which the TF returns to the target via looping. This is underlined by the fact that the peak is close to the time scale τ_{init} which for this parameter set is $\tau_{\text{init}} \approx 1.4 \times 10^3$. In fact, the peak is at slightly larger times since the loop initiation time is just a part of the whole time needed for forming a loop.

We have to comment on the reduced size of the peak for fast rebinding events. This is due to the way our algorithm treats quick rebinding in a looped state. While these events were counted explicitly in the system without looping, in a looped state only the times when the particle dissociates from the target region are counted. This already takes into account the biological interpretation mentioned above. Effectively, looping introduces a new time scale for rebinding in comparison to long-lasting events involving complete dissociation from the target region. This re-establishes the experimental findings of Paul J. Choi et al. that in a system with looping gene expression is regulated on different time

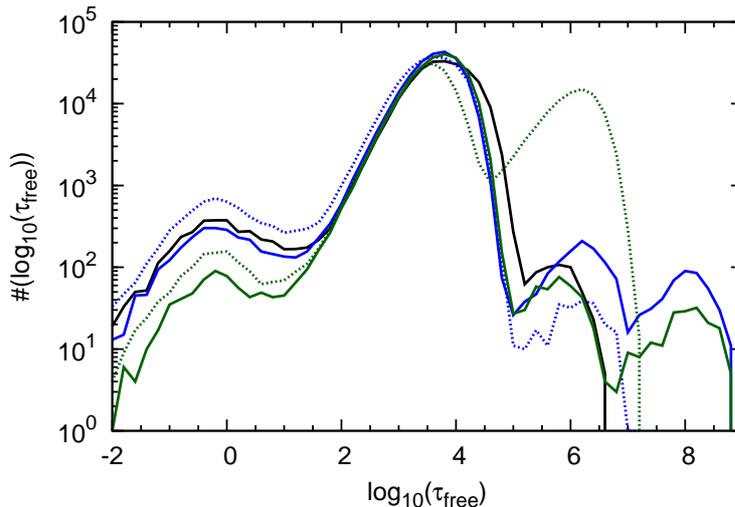


Figure 5.9: Distribution of periods in which the main operator is not bound by the repressor for different blocker configurations. In all configurations looping is possible. Black: only $O2$ is accessible. Blue: only $O3$ is accessible. Green: both $O2$ and $O3$ are accessible. Dashed lines: $O3$ is within the target region, full lines: $O3$ is outside the target region. In all cases $\alpha = 0.6$.

scales [18]. More explicitly, they found small expression bursts occur when a loop opens shortly at the main operator, while large bursts happen when a complete dissociation from both operators occurs.

This interpretation is further substantiated by looking at the dashed line which shows results from simulations in which single bulk excursions on average last ten times longer as well as looping does. While the position of the first peak is unaffected by this change, the other two peaks are shifted to the right as expected.

5.4.3 Blocker conformation effects

Finally, we study the influence of different blocker configurations which all allow looping. The black line in Fig. 5.9 corresponds to the case where only $O2$ is accessible and is equal to the blue line in the previous Fig. 5.8. Conversely, the blue lines correspond to the situation when only the weaker $O3$ can be bound. The mixed case, when both auxiliary operators can be bound are shown in green. In both cases when $O3$ is accessible, it is further distinguished whether or not $O3$ lies within the target region. If so, it is represented by a dashed line in Fig. 5.9, if not by a full line.

We observe that all curves have at least three peaks, and two of the curves have four. The three peaks which are always present have the same interpretation as before. The fourth peak only shows up if the auxiliary operator $O3$ is present, but not in the target region. We interpret these very persistent events as repeated binding to the auxiliary operator, but unlike in the case when it lies in the target region after leaving $O3$ the particle cannot slide towards the main target since there is a roadblock in between.

5.5 Summary

The model presented in this chapter is meant to describe the association between the lac repressor and its main operator $O1$ in a living *E. coli* cell. Unlike the treatment in previous chapters we assumed that non-specifically bound proteins act as roadblocks and partition the one-dimensional search space on DNA into a set of intervals. In the first part of this chapter we studied the last search phase when a TF has just landed at a random position in the target region, i.e. between the two roadblocks which are closest to the main operator.

This was done in terms of a numerical simulation and a simplified theoretical model. In both cases, the protein could be present in the two conformations which had been introduced before, the search state and the recognition state. To get more realistic results, in this chapter the interconversion rates between these states at all positions—in particular at the target—depended on the underlying nucleotide sequence. For the probability to detect the target before dissociation, the numerical results agreed very well with the theoretical model in which all non-target sites were assumed to be equivalent. However, for the calculation of the conditional search times, this assumption breaks down if binding to an auxiliary operator is possible. Such calculations of the target detection probability for a starting position which is already close to the target, are also important in light of the co-localisation effect which was discussed in section 2.4 [114, 115, 212].

In the simplified setting focusing exclusively on what happens in the target region such an auxiliary operator behaves as a trap for the searcher. In the chapter's second part it was established that a TF which is able to form DNA loops by binding simultaneously to two operators can turn this liability into an asset. For the loop-forming tetrameric repressor the auxiliary operator is a short-cut from which it can quickly bind the main operator. Depending on which auxiliary operators were accessible and whether direct sliding from one operator to another was possible, we obtained a rich behaviour on many different time scales.

When studying a histogram of the biologically relevant time spans during which the main operator is free of repressor, the number of peaks follows a simple rule: two peaks are always present, representing quick returns to the target without dissociation and long-lasting return events which involve several dissociations. Whenever an auxiliary operator is present, a third peak appears indicating rebinding events by looping. For reasonable parameter values this happens on a scale shorter than the event involving complete dissociation. The peak of time-consuming return splits into two when the weaker auxiliary operator is not in the same interval between roadblocks as the main operator. In general, this model underlines that looping is a measure of the cell how a small number of TFs can efficiently locate a main operator without being present in large copy numbers, since this would contribute to crowding effects [141].

6 Discussion and outlook

How seemingly simple living organisms like bacteria are able to survive in adverse conditions is a fascinating question in biology. Partly, this is due to the efficient way in which they make use of their genes. This gene expression can—to some extent—be described with physical theories. By default, the focus of experiments which concerned the association reaction of transcription factors with their operator sequences on DNA and which were performed in the last century, was on *in vitro* studies. The facilitated diffusion model introduced by Berg, Winter, von Hippel and co-workers was able to explain the observed high association rates as a result of a beneficial combination of three-dimensional search phases in the bulk solution with one-dimensional sliding along DNA. The interest of theoreticians increased in recent years when more and more experiments with single-molecule resolution became possible, in some cases even in living bacterial cells.

Obviously, it is by far more important to describe this biological search process in a living cell than to explain high association rates in a petri dish. However, the crowded interior of a living cell is a completely different search space than the one encountered in a dilute *in vitro* assay. This observation and the fact that an increasing number of microscopic parameters is known constitute a challenge to adapt the long-established theoretical models which were described in the second chapter of this work to these new insights.

The approach presented in chapter 3 therefore combined the original facilitated diffusion model with the common concept that the searching particle is present in two conformations. These two conformations enable the particle to translocate quickly along DNA while retaining the ability to bind strongly to the target sequence. Importantly, our result reproduces the early measured operator association rates and as usual, the rate at which the target is detected depends non-monotonically on the non-specific binding affinity. This underlines that in general only an appropriately chosen composition of both search mechanisms is an efficient strategy. Going beyond simpler models, our model shows that for relevant parameter values more search time should be spent sliding than in the bulk solution—as recently found in experiments.

Chapter 4 directly approached the situation in a living cell. A simple model of the bacterial genome was taken from the literature and coupled to a compound of microscopic facilitated diffusion models. Again, the result showed an excellent agreement with recent experimental *in vivo* findings. Interestingly, within our model the search time did not depend strongly on the exact parameter values showing that the operator association reaction occurs at nearly optimal search conditions. Yet, the search could not be sped up much by choosing a starting position for the particle within the subunit of the genome which hosts the target sequence. However, the subunit in this setting represented a rather large volume, motivating an even more local consideration as presented in chapter 5.

Here the probability was calculated to detect the target before dissociation from DNA

occurs when the search starts on a non-specific binding site which is already very close to the target. An impressive accord of a simplified theoretical model with numerical simulations involving the real nucleotide sequence of a common *E. coli* strain was observed. Advancing from probabilities to search *times*, the presence and accessibility of auxiliary operators becomes increasingly important, in particular when the ability of the searcher to simultaneously bind to two operators is taken into account.

Roughly, the fraction of numerical results increased from chapter to chapter at the cost of the fraction of analytical results. However, as seen in the last chapter theoretical models still have quite some descriptive power when it is assured that the assumptions they rely on are meaningful. Accordingly, each chapter presented its own model which highlighted specific aspects of the search process.

The general consensus in the field seems to be that invoking the facilitated diffusion mechanism of sliding and relocations via the three-dimensional space makes sense for searching particles which are present in rather low copy numbers [125]. Other proteins which are non-specifically bound can be perceived as roadblocks for the sliding motion of the particle [141]. While roadblock *a priori* is a term with a rather bad connotation, under certain circumstances they can have a beneficial influence, for example when they block non-specific regions on DNA where searching would be a waste of time [122]. However, they should not be too close to the target in order not to cut off too much of the “antenna” of non-specific sites near the operator. There are some hints that this dual role is indeed observed in some eukaryotes [122].

In many instances it is advisable to take a down-to-earth approach. From the observation that sliding along DNA is an experimental reality and induces a larger effective target size, one should not automatically infer that the facilitated diffusion mechanism enables association rates above the diffusion limit. Rather the conclusion should be that DNA-binding proteins make the best of their general affinity for non-specific DNA.

In general, no premature conclusions should be drawn. Apparently, nature often preferred to choose a solution to a problem which at first sight does not seem to be the easiest way. For example, in the *lac* system one might expect that increasing the concentration of searching particles and maximising the binding affinity for the target is the straightforward way to achieve fast and reliable gene regulation [44]. However, the natural main operator is not the strongest possible one and the concentration of searching repressors is rather low. Instead the repressor is usually present in its tetrameric form and forms DNA loops between operators whose specificities are not optimised [44]. Thus, it is not the global concentration of TFs which is maximised, but only the *local* concentration close to the target sequence via looping [49].

While the models presented in this work successfully describe many of the experimental findings, they are obviously not yet the end of the line. Many more factors or boundary conditions can play a role, since for example, a cell has only finite resources. Additionally, for a complete description of gene regulation the rates at which mRNA molecules and proteins are degraded have to be taken into account and some proteins need to undergo dimerisation before they can perform certain tasks. Anomalous transport mechanisms were neglected so far, even though some proteins in crowded solutions show such a behaviour [93]. Furthermore, we usually modelled the behaviour of a single searcher and therefore no dependence on the concentration of searchers was discussed. However, for example the results presented in the last chapter can be used as input for thermodynamic

models which are generally very successful in describing the gene expression output and which automatically take concentration effects into account [61]. Besides, one has to note that while this work focused on the association of TFs with their target sequence, the related question how accurate or noisy the expression of a gene under regulation of such a TF is, represents a stand-alone research topic [214, 215]. A more or less direct extension of the models presented herein into this direction would be to calculate higher moments of the first passage time distribution to the target.

There are hints that the paradigmatic lac repressor which also in this work was taken again and again as an example of use might not be a typical representative of DNA-binding proteins [125]. After all, there might be more important tasks to be carried out than to make sure that glucose is metabolised instead of lactose. However, it is improbable that the lac repressor should be the only protein to make direct use of its non-specific affinity for DNA. Therefore, it is highly desirable to have more detailed experimental data also for other DNA-binding proteins. Based on nature's ability to surprise scientists, one can be sure that in the future many more interesting results in theory, simulations and experiments will be found.

Acknowledgment

This work would not have been possible without the help of several people whom I had the pleasure to meet in the last few years. Foremost, I would like to thank Prof. Ralf Metzler for giving me the opportunity to work on this PhD project in his group and giving me support and technical advice throughout this time. I admire his broad spectrum of knowledge in theoretical physics and him doing his magic on scientific texts, but I also enjoyed talking about our home in the “south-west”.

Since most of this work is devoted to the calculation of **mean first passage times**, I would like to thank people I met in **MFPT**, i.e. **M**ünchen, **F**reiburg, **P**otsdam, and **T**ampere.

München I gladly recollect the early days in Munich. It was a very nice time with Jochen Kursawe and Mahsa Vahabi and in particular with my room-mates Leila Esmaeili Sereshki and Johannes H. P. Schulz. A special thanks goes out to Vladimir V. Palyulin. His ceaseless helpfulness especially as a host during the phase transition Munich-Potsdam is greatly appreciated.

Tampere I thank the Finnish part of our group, namely Otto Pulkkinen and Jae-Hyung Jeon for the warm welcome during my two stays in the far north and the nice time we spent when they came to Germany.

Potsdam I enjoyed very much the familiar atmosphere at the Institute of Physics & Astronomy of Potsdam University in general, and in particular in our group. Namely, I would like to thank all members of our group: Andrey G. Cherstvy, Fred Feudel, Igor Goychuk, Maria & Richard Schwarzl and the Fraunhofer canteen regulars Henning Krüseemann, Jaehoh Shin and Surya Kanta Ghosh. I thank our guest Prof. Michael Andersen Lomholt for stimulating discussions.

I am particularly thankful to Aljaž Godec: it was a pleasure to work on the common project, but also to discuss physics and other aspects of life over litres of coffee. Furthermore, I like to thank my room-mates Darius Schweinoch and Azamat Yeldesbay for making all the small breaks such pleasant distractions.

Freiburg Even with the technical help of all the people mentioned above, I would not have been able to fulfil this work without the continuous support of my parents Gabriele & Michael Bauer, my sister Friederike Bauer, as well as my “in-laws” Rositha Maier-Eisenbach and Peter Eisenbach. Most importantly, I want to express my endless gratitude towards Stephanie Eisenbach. Firstly, for proofreading the first sketch of this thesis, but mainly for all the things that are hard to put into words. Thank you for being you!

Bibliography

- [1] I. Burnet, *Platonis Opera*, (Oxford University Press, 1979), sixteenth impression.
- [2] W. Johannsen, *Naturwissenschaften* **5**, 389 (1917).
- [3] R. Dahm, *Dev. Biol.* **278**, 274 (2005).
- [4] G. Mendel, *Verhandlungen des naturforschenden Vereines in Brünn* **4**, 3 (1866).
- [5] I. Miko, *Nature Education* **1**, 134 (2008).
- [6] J. D. Watson and F. H. C. Crick, *Nature* **171**, 737 (1953).
- [7] W. S. Sutton, *Biol. Bull.* **4**, 24 (1902).
- [8] C. O'Connor, *Nature Education* **1**, 105 (2008).
- [9] F. Griffith, *J. Hyg. (Lond.)* **27**, 113 (1928).
- [10] O. T. Avery, C. M. MacLeod, and M. McCarty, *J. Exp. Med.* **79**, 137 (1944).
- [11] E. Schrödinger, *What is life?* (Cambridge University Press, 1944).
- [12] A. D. Hershey and M. Chase, *J. Gen. Physiol.* **36**, 39 (1952).
- [13] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell* (Garland Science, Taylor & Francis Group, 2007).
- [14] S. E. Luria and M. Delbrück, *Genetics* **28**, 491 (1943).
- [15] G. W. Beadle and E. L. Tatum, *Proc. Natl. Acad. Sci. U. S. A.* **27**, 499 (1941).
- [16] M. Ptashne, *A genetic switch - Third Edition - Phage Lambda Revisited* (Cold Spring Harbor, New York, 2004).
- [17] A. Novick and M. Weiner, *Proc. Natl. Acad. Sci. U. S. A.* **43**, 553 (1957).
- [18] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie, *Science* **322**, 442 (2008).
- [19] R. F. Bruinsma, *Physica A* **313**, 211 (2002).
- [20] F. H. C. Crick, *Nature* **227**, 561 (1970).
- [21] J. Romanuka *et al.*, *J. Mol. Biol.* **390**, 478 (2009).
- [22] H. M. Berman *et al.*, *Nucleic Acids Res.* **28**, 235 (2000).

- [23] J. Mathé, A. Aksimentiev, D. R. Nelson, K. Schulten, and A. Meller, Proc. Natl. Acad. Sci. U. S. A. **102**, 12377 (2005).
- [24] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie, Science **311**, 1600 (2006).
- [25] G.-W. Li and X. S. Xie, Nature **475**, 308 (2011).
- [26] F. H. C. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin, Nature **192**, 1227 (1961).
- [27] M. W. Nirenberg and J. H. Matthaei, Proc. Natl. Acad. Sci. U. S. A. **47**, 1588 (1961).
- [28] M. W. Nirenberg *et al.*, Proc. Natl. Acad. Sci. U. S. A. **53**, 1161 (1965).
- [29] N. N. Ivanova *et al.*, Science **344**, 909 (2014).
- [30] I. M. Keseler *et al.*, Nucleic Acids Res. **41**, D605 (2013).
- [31] T. M. Earnest, E. Roberts, M. Assaf, K. Dahmen, and Z. Luthey-Schulten, Phys. Biol. **10**, 026002 (2013).
- [32] P. Hammar, *lac of Time - Transcription Factor Kinetics in Living Cells*, PhD thesis, Acta Universitatis Upsaliensis, 2013.
- [33] K. J. Andrews and E. Lin, J. Bacteriol. **128**, 510 (1976).
- [34] A. Lewendon, J. Ellis, and W. V. Shaw, J. Biol. Chem. **270**, 26326 (1995).
- [35] M. Demerec, E. A. Adelberg, A. Clark, and P. E. Hartman, Genetics **54**, 61 (1966).
- [36] W. Gilbert and B. Müller-Hill, Proc. Natl. Acad. Sci. U. S. A. **58**, 2415 (1967).
- [37] W. Gilbert and B. Müller-Hill, Proc. Natl. Acad. Sci. U. S. A. **56**, 1891 (1966).
- [38] M. Ptashne, *A Genetic Switch: Gene Control and Phage Lambda* (Blackwell Science Ltd.) 2nd revised edition (1992).
- [39] J. Elf, G.-W. Li, and X. S. Xie, Science **316**, 1191 (2007).
- [40] T. Kuhlman, Z. Zhang, M. H. Saier, and T. Hwa, Proc. Natl. Acad. Sci. U. S. A. **104**, 6043 (2007).
- [41] J. Monod, A. M. Pappenheimer, and G. Cohen-Bazire, Biochim. Biophys. Acta **9**, 648 (1952).
- [42] J. R. Sadler, H. Sasmor, and J. L. Betz, Proc. Natl. Acad. Sci. U. S. A. **80**, 6785 (1983).
- [43] W. S. Reznikoff, R. B. Winter, and C. K. Hurley, Proc. Natl. Acad. Sci. U. S. A. **71**, 2314 (1974).
- [44] S. Oehler, E. R. Eismann, H. Krämer, and B. Müller-Hill, EMBO J. **9**, 973 (1990).

-
- [45] E. Villa, A. Balaeff, and K. Schulten, Proc. Natl. Acad. Sci. U. S. A. **102**, 6783 (2005).
- [46] J. Kania and D. T. Brown, Proc. Natl. Acad. Sci. U. S. A. **73**, 3529 (1976).
- [47] D. Swigon and W. K. Olson, Int. J. Non Linear Mech. **43**, 1082 (2008).
- [48] R. Schleif, Annu. Rev. Biochem. **61**, 199 (1992).
- [49] B. Müller-Hill, Mol. Microbiol. **29**, 13 (1998).
- [50] R. J. Ellis, Trends Biochem. Sci. **26**, 597 (2001).
- [51] A. P. Minton, J. Biol. Chem. **276**, 10577 (2001).
- [52] E. Barkai, Y. Garini, and R. Metzler, Physics Today **65**, 29 (2012).
- [53] J. Kania and B. Müller-Hill, Eur. J. Biochem. **79**, 381 (1977).
- [54] D. Pettijohn and T. Kamiya, J. Mol. Biol. **29**, 275 (1967).
- [55] A. D. Riggs, S. Bourgeois, and M. Cohn, J. Mol. Biol. **53**, 401 (1970).
- [56] P. H. von Hippel and O. G. Berg, Proc. Natl. Acad. Sci. U. S. A. **83**, 1608 (1986).
- [57] K. Loth, M. Gnida, J. Romanuka, R. Kaptein, and R. Boelens, J. Biomol. NMR **56**, 41 (2013).
- [58] P. L. DeHaseth, T. M. Lohman, and M. T. Record Jr., Biochemistry **16**, 4783 (1977).
- [59] R. B. Winter and P. H. von Hippel, Biochemistry **20**, 6948 (1981).
- [60] P. H. von Hippel, A. Revzin, C. A. Gross, and A. C. Wang, Proc. Natl. Acad. Sci. U. S. A. **71**, 4808 (1974).
- [61] J. M. G. Vilar and L. Saiz, ACS Synth. Biol. **2**, 576 (2013).
- [62] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. U. S. A. **99**, 12015 (2002).
- [63] G. K. Ackers, A. D. Johnson, and M. A. Shea, Proc. Natl. Acad. Sci. U. S. A. **79**, 1129 (1982).
- [64] A. Bakk and R. Metzler, FEBS Lett. **563**, 66 (2004).
- [65] H. G. Garcia *et al.*, Cell Rep. **2**, 150 (2012).
- [66] M. von Smoluchowski, Phys. Z. **17**, 557 (1916).
- [67] P. H. von Hippel and O. G. Berg, J. Biol. Chem. **264**, 675 (1989).
- [68] A. Szabo, K. Schulten, and Z. Schulten, J. Chem. Phys. **72**, 4350 (1980).

- [69] S.-Y. Lin and A. D. Riggs, *J. Mol. Biol.* **72**, 671 (1972).
- [70] S. Bourgeois and A. D. Riggs, *Biochem. Biophys. Res. Commun.* **38**, 348 (1970).
- [71] R. Fickert and B. Müller-Hill, *J. Mol. Biol.* **226**, 59 (1992).
- [72] S. E. Halford, *Biochem. Soc. Trans.* **37**, 343 (2009).
- [73] A.-M. Florescu and M. Joyeux, *J. Phys. Chem. A* **114**, 9662 (2010).
- [74] P. Hänggi, P. Talkner, and M. Borkovec, *Rev. Mod. Phys.* **62**, 251 (1990).
- [75] R. B. Winter, O. G. Berg, and P. H. von Hippel, *Biochemistry* **20**, 6961 (1981).
- [76] D. Stigter, *Biopolymers* **16**, 1435 (1977).
- [77] O. G. Berg and C. Blomberg, *Biophys. Chem.* **8**, 271 (1978).
- [78] G. Pólya, *Math. Ann.* **84**, 149 (1921).
- [79] M. A. McCloskey and M. Poo, *J. Cell Biol.* **102**, 88 (1986).
- [80] G. Adam and M. Delbrück, *Reduction of Dimensionality in Biological Diffusion Processes in Structural Chemistry and Molecular Biology* (W. H. Freeman & Co., 1968).
- [81] F. Wang *et al.*, *Nat. Struct. Mol. Biol.* **20**, 174 (2013).
- [82] P. H. Richter and M. Eigen, *Biophys. Chem.* **2**, 255 (1974).
- [83] R. Schraner and P. Richter, *Biophys. Chem.* **8**, 135 (1978).
- [84] O. G. Berg and C. Blomberg, *Biophys. Chem.* **4**, 367 (1976).
- [85] O. G. Berg, R. B. Winter, and P. H. von Hippel, *Biochemistry* **20**, 6929 (1981).
- [86] O. G. Berg and C. Blomberg, *Biophys. Chem.* **7**, 33 (1977).
- [87] O. G. Berg and C. Blomberg, *Biophys. Chem.* **9**, 415 (1979).
- [88] A. Veksler and A. B. Kolomeisky, *J. Phys. Chem. B* **117**, 12695 (2013).
- [89] V. Dahirel, F. Paillusson, M. Jardat, M. Barbi, and J.-M. Victor, *Phys. Rev. Lett.* **102**, 228101 (2009).
- [90] O. Bénichou, Y. Kafri, M. Sheinman, and R. Voituriez, *Phys. Rev. Lett.* **103**, 138102 (2009).
- [91] N. Shimamoto, *J. Biol. Chem.* **274**, 15293 (1999).
- [92] M. Hsieh and M. Brenowitz, *J. Biol. Chem.* **272**, 22092 (1997).
- [93] I. Golding and E. C. Cox, *Phys. Rev. Lett.* **96**, 098102 (2006).

-
- [94] M. B. Elowitz, M. G. Surette, P.-E. Wolf, J. B. Stock, and S. Leibler, *J. Bacteriol.* **181**, 197 (1999).
- [95] P. Montero Llopis *et al.*, *Nature* **466**, 77 (2010).
- [96] D. M. Gowers, G. G. Wilson, and S. E. Halford, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15883 (2005).
- [97] I. M. Sokolov, R. Metzler, K. Pant, and M. C. Williams, *Biophys. J.* **89**, 895 (2005).
- [98] Y. M. Wang, R. H. Austin, and E. C. Cox, *Phys. Rev. Lett.* **97**, 048302 (2006).
- [99] I. Bonnet *et al.*, *Nucleic Acids Res.* **36**, 4118 (2008).
- [100] P. C. Blainey *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5752 (2006).
- [101] P. C. Blainey *et al.*, *Nat. Struct. Mol. Biol.* **16**, 1224 (2009).
- [102] A. Tafvizi, F. Huang, A. R. Fersht, L. A. Mirny, and A. M. van Oijen, *Proc. Natl. Acad. Sci. U. S. A.* **108**, 563 (2011).
- [103] J. S. Leith *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16552 (2012).
- [104] P. Hammar *et al.*, *Science* **336**, 1595 (2012).
- [105] A. Marcovitz and Y. Levy, *Biophys. J.* **104**, 2042 (2013).
- [106] P. Hammar *et al.*, *Nat. Genet.* **46**, 405 (2014).
- [107] Y. Taniguchi *et al.*, *Science* **329**, 533 (2010).
- [108] M. Razo-Mejia *et al.*, *Phys. Biol.* **11**, 026005 (2014).
- [109] J. B. Kinney, A. Murugan, J. C. G. Callan, and E. C. Cox, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9158 (2010).
- [110] J. Q. Boedicker, H. G. Garcia, S. Johnson, and R. Phillips, *Phys. Biol.* **10**, 066005 (2013).
- [111] N. A. Becker, J. P. Peters, T. A. Lionberger, and L. J. Maher, *Nucleic Acids Res.* **41**, 156 (2013).
- [112] F. Képès, *J. Mol. Biol.* **340**, 957 (2004).
- [113] P. Warren and P. R. ten Wolde, *J. Mol. Biol.* **342**, 1379 (2004).
- [114] G. Kolesov, Z. Wunderlich, O. N. Laikova, M. S. Gelfand, and L. A. Mirny, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13948 (2007).
- [115] Z. Wunderlich and L. A. Mirny, *Nucleic Acids Res.* **36**, 3570 (2008).
- [116] S. C. Janga, H. Salgado, and A. Martínez-Antonio, *Nucleic Acids Res.* **37**, 3680 (2009).

- [117] I. Junier, M. Olivier, and F. Képès, PLoS Comput. Biol. **6**, e1000678 (2010).
- [118] M. Fritsche, S. Li, D. W. Heermann, and P. A. Wiggins, Nucleic Acids Res. **40**, 972 (2012).
- [119] T. E. Kuhlman and E. C. Cox, Mol. Syst. Biol. **8**, 610 (2012).
- [120] O. Pulkkinen and R. Metzler, Phys. Rev. Lett. **110**, 198101 (2013).
- [121] T. E. Kuhlman and E. C. Cox, Phys. Rev. E **88**, 022701 (2013).
- [122] L. A. Mirny *et al.*, J. Phys. A **42**, 434013 (2009).
- [123] A. B. Kolomeisky, Phys. Chem. Chem. Phys. **13**, 2088 (2011).
- [124] M. Sheinman, O. Bénichou, Y. Kafri, and R. Voituriez, Rep. Prog. Phys. **75**, 026601 (2012).
- [125] S. Redding and E. C. Greene, Chem. Phys. Lett. **570**, 1 (2013).
- [126] M. Slutsky and L. A. Mirny, Biophys. J. **87**, 4021 (2004).
- [127] M. Coppey, O. Bénichou, R. Voituriez, and M. Moreau, Biophys. J. **87**, 1640 (2004).
- [128] I. Eliazar, T. Koren, and J. Klafter, J. Phys.: Condens. Matter **19**, 065140 (2007).
- [129] Y. Meroz, I. Eliazar, and J. Klafter, J. Phys. A **42**, 434012 (2009).
- [130] C. Loverdo *et al.*, Phys. Rev. Lett. **102**, 188101 (2009).
- [131] T. Hu and B. I. Shklovskii, Phys. Rev. E **74**, 021903 (2006).
- [132] T. Hu and B. I. Shklovskii, Phys. Rev. E **76**, 051909 (2007).
- [133] M. A. Lomholt, T. Ambjörnsson, and R. Metzler, Phys. Rev. Lett. **95**, 260603 (2005).
- [134] T. Hu, A. Y. Grosberg, and B. I. Shklovskii, Biophys. J. **90**, 2731 (2006).
- [135] M. A. Díaz de la Rosa, E. F. Koslover, P. J. Mulligan, and A. J. Spakowitz, Biophys. J. **98**, 2943 (2010).
- [136] E. F. Koslover, M. A. Díaz de la Rosa, and A. J. Spakowitz, Biophys. J. **101**, 856 (2011).
- [137] B. van den Broek, M. A. Lomholt, S.-M. J. Kalisch, R. Metzler, and G. J. L. Wuite, Proc. Natl. Acad. Sci. U. S. A. **105**, 15738 (2008).
- [138] M. A. Lomholt, B. van den Broek, S.-M. J. Kalisch, G. J. L. Wuite, and R. Metzler, Proc. Natl. Acad. Sci. U. S. A. **106**, 8204 (2009).
- [139] T. Schötz, R. A. Neher, and U. Gerland, Phys. Rev. E **84**, 051911 (2011).

-
- [140] H. Flyvbjerg, S. A. Keatch, and D. T. F. Dryden, *Nucleic Acids Res.* **34**, 2550 (2006).
- [141] G.-W. Li, O. G. Berg, and J. Elf, *Nat. Phys.* **5**, 294 (2009).
- [142] M. Tabaka, T. Kalwarczyk, and R. Holyst, *Nucleic Acids Res.* **42**, 727 (2014).
- [143] K. V. Klenin, H. Merlitz, J. Langowski, and C.-X. Wu, *Phys. Rev. Lett.* **96**, 018104 (2006).
- [144] H. Merlitz, K. Klenin, C.-X. Wu, and J. Langowski, *J. Chem. Phys.* **124**, 134908 (2006).
- [145] A. G. Cherstvy, A. B. Kolomeisky, and A. A. Kornyshev, *J. Phys. Chem. B* **112**, 4741 (2008).
- [146] O. G. Berg and M. Ehrenberg, *Biophys. Chem.* **15**, 41 (1982).
- [147] H.-X. Zhou and A. Szabo, *Phys. Rev. Lett.* **93**, 178101 (2004).
- [148] H.-X. Zhou, *Biophys. J.* **88**, 1608 (2005).
- [149] B. P. Belotserkovskii and D. A. Zarling, *J. Theor. Biol.* **226**, 195 (2004).
- [150] K. Park, T. Kim, and H. Kim, *Bull. Korean Chem. Soc.* **33**, 971 (2012).
- [151] G. Malherbe and D. Holcman, *Phys. Lett. A* **374**, 466 (2010).
- [152] N. R. Zabet, *BMC Syst. Biol.* **6**, 121 (2012).
- [153] N. R. Zabet and B. Adryan, *Mol. Biosyst.* **8**, 2815 (2012).
- [154] N. R. Zabet and B. Adryan, *Bioinformatics* **28**, 1287 (2012).
- [155] N. R. Zabet and B. Adryan, *Bioinformatics* **28**, 1517 (2012).
- [156] N. R. Zabet and B. Adryan, *Frontiers in genetics* **4**, 197 (2013).
- [157] N. R. Zabet, R. Foy, and B. Adryan, *PLOS ONE* **8**, e73714 (2013).
- [158] D. Ezer, N. R. Zabet, and B. Adryan, *Nucleic Acids Res.* **42**, 4196 (2014).
- [159] A.-M. Florescu and M. Joyeux, *J. Chem. Phys.* **130**, 015103 (2009).
- [160] A.-M. Florescu and M. Joyeux, *J. Chem. Phys.* **131**, 105102 (2009).
- [161] G. Foffano, D. Marenduzzo, and E. Orlandini, *Phys. Rev. E* **85**, 021919 (2012).
- [162] C. A. Brackley, M. E. Cates, and D. Marenduzzo, *Phys. Rev. Lett.* **111**, 108101 (2013).
- [163] B. Richey *et al.*, *J. Biol. Chem.* **262**, 7157 (1987).
- [164] Y. von Hansen, R. R. Netz, and M. Hinczewski, *J. Chem. Phys.* **132**, 135103 (2010).

- [165] R. Zwanzig, Proc. Natl. Acad. Sci. U. S. A. **85**, 2029 (1988).
- [166] C. G. Kalodimos *et al.*, Science **305**, 386 (2004).
- [167] L. Hu, A. Y. Grosberg, and R. Bruinsma, Biophys. J. **95**, 1151 (2008).
- [168] J. Reingruber and D. Holcman, J. Phys.: Condens. Matter **22**, 065103 (2010).
- [169] J. Reingruber and D. Holcman, Phys. Rev. E **84**, 020901 (2011).
- [170] R. Murugan, J. Phys. A **43**, 195003 (2010).
- [171] H.-X. Zhou, Proc. Natl. Acad. Sci. U. S. A. **108**, 8651 (2011).
- [172] S. Yu, S. Wang, and R. G. Larson, J. Biol. Phys. **39**, 565 (2013).
- [173] S. E. Halford and J. F. Marko, Nucleic Acids Res. **32**, 3040 (2004).
- [174] M. Bauer and R. Metzler, Biophys. J. **102**, 2321 (2012).
- [175] A. V. Chechkin, I. M. Zaid, M. A. Lomholt, I. M. Sokolov, and R. Metzler, J. Chem. Phys. **134**, 204116 (2011).
- [176] M. Kampmann, J. Biol. Chem. **279**, 38715 (2004).
- [177] A. Jeltsch and A. Pingoud, Biochemistry **37**, 2160 (1998).
- [178] A. V. Chechkin, I. M. Zaid, M. A. Lomholt, I. M. Sokolov, and R. Metzler, Phys. Rev. E **79**, 040105 (2009).
- [179] F. Y.-H. Wu, P. Bandyopadhyay, and C.-W. Wu, J. Mol. Biol. **100**, 459 (1976).
- [180] M. Bauer, A. Godec, and R. Metzler, Phys. Chem. Chem. Phys. **16**, 6118 (2014).
- [181] A. Godec, M. Bauer, and R. Metzler, arXiv preprint arXiv:1403.3910 (2014).
- [182] M. Bauer and R. Metzler, PLOS ONE **8**, e53956 (2013).
- [183] E. P. C. Rocha, Annu. Rev. Genet. **42**, 211 (2008).
- [184] J. Kim *et al.*, Nucleic Acids Res. **32**, 1982 (2004).
- [185] L. Postow, C. D. Hardy, J. Arsuaga, and N. R. Cozzarelli, Genes Dev. **18**, 1766 (2004).
- [186] T. Romantsov, I. Fishov, and O. Krichevsky, Biophys. J. **92**, 2875 (2007).
- [187] M. A. Umbarger *et al.*, Mol. Cell **44**, 252 (2011).
- [188] P. H. Viollier *et al.*, Proc. Natl. Acad. Sci. U. S. A. **101**, 9257 (2004).
- [189] S. Jun and A. Wright, Nat. Rev. Microbiol. **8**, 600 (2010).
- [190] Y. Jung *et al.*, Soft Matter **8**, 2095 (2012).

-
- [191] M. Buenemann and P. Lenz, PLOS ONE **5**, e13806 (2010).
- [192] N. Madras, A. Orlicsky, and L. A. Shepp, J. Stat. Phys. **58**, 159 (1990).
- [193] O. Bénichou, C. Chevalier, J. Klafter, B. Meyer, and R. Voituriez, Nat. Chem. **2**, 472 (2010).
- [194] B. Meyer, C. Chevalier, R. Voituriez, and O. Bénichou, Phys. Rev. E **83**, 051116 (2011).
- [195] G. Guigas and M. Weiss, Biophys. J. **94**, 90 (2008).
- [196] L. E. Sereshki, M. A. Lomholt, and R. Metzler, Europhys. Lett. **97**, 20008 (2012).
- [197] E. S. Rasmussen, *Protein Diffusion on Heterogeneous DNA*, Bachelor thesis, University of Southern Denmark, 2013.
- [198] J. D. McGhee and P. H. von Hippel, J. Mol. Biol. **86**, 469 (1974).
- [199] L. Hu, A. Y. Grosberg, and R. Bruinsma, J. Phys. A **42**, 434011 (2009).
- [200] D. T. Gillespie, J. Comput. Phys. **22**, 403 (1976).
- [201] M. Barbi, C. Place, V. Popkov, and M. Salerno, J. Biol. Phys. **30**, 203 (2004).
- [202] W. W. Wasserman and A. Sandelin, Nat. Rev. Genet. **5**, 276 (2004).
- [203] J. M. Vilar, Biophys. J. **99**, 2408 (2010).
- [204] G. D. Stormo and D. S. Fields Trends Biochem. Sci. **23**, 109 (1998).
- [205] S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill, EMBO J. **13**, 3348 (1994).
- [206] H. G. Garcia and R. Phillips, Proc. Natl. Acad. Sci. U. S. A. **108**, 12173 (2011).
- [207] A. B. Kolomeisky and A. Veksler, J. Chem. Phys. **136**, 125101 (2012).
- [208] A. Esadze and J. Iwahara, J. Mol. Biol. **426**, 230 (2014).
- [209] K. Spendier and V. M. Kenkre, J. Phys. Chem. B **117**, 15639 (2013).
- [210] R. M. Pearlstein, J. Chem. Phys. **56**, 2431 (1972).
- [211] K. Lakatos-Lindenberg, R. P. Hemenger, and R. M. Pearlstein, J. Chem. Phys. **56**, 4852 (1972).
- [212] O. Bénichou, C. Loverdo, and R. Voituriez, Europhys. Lett. **84**, 38003 (2008).
- [213] D. G. Priest *et al.*, Proc. Natl. Acad. Sci. U. S. A. **111**, 349 (2014).
- [214] J. S. van Zon, M. J. Morelli, S. Tănase-Nicola, and P. R. ten Wolde, Biophys. J. **91**, 4350 (2006).
- [215] G. Tkačik and W. Bialek, Phys. Rev. E **79**, 051901 (2009).