uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Nazanin Samadifard
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Physics)
GRADE / DEGREE

Department of Physics
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Facilitated Diffusion of Proteins or DNA

TITRE DE LA THÈSE / TITLE OF THESIS

J. Harden
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

B. Joos

A. Pelling

G. Oarham

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Facilitated diffusion of proteins on DNA

by

Nazanin Samadifard

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the M.Sc. degree in
Physics

Department of Physics
Faculty of Science
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-65549-8
Our file Notre référence
ISBN: 978-0-494-65549-8

Table of Contents

List of Figures

List of Tables

Acronyms

ADP	Adenosine Diphosphate
ATP	Adenosine Triphosphate
BD	Brownian dynamics simulation
bp	base pair
DNA	Deoxyribonucleic acid
FENE	finitely extenteded elastic
FPT	first passage time
HOgg1	Human oxoguanine DNA glycosylase
HoxD9	Engrailed Home domain
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator
LJ	Lennard Jones
MFPT	mean first passage time
MSD	mean square displacement
PBC	periodic boundary condition
RNA	Ribonucleic acid
TIRFM	total internal reflection orescent microscopy
Vitro	latin: within the glass
Vivo	latin for "within the living"
3D	three-dimensional

Abstract

The idea that non-specific DNA-binding proteins are capable of finding their cognate sites on DNA much faster than the time calculated via three-dimensional diffusion theory has been proposed in recent years as an important mechanism for the activity of DNA enzymes and transcription factors. The goal of this study was to investigate, using Brownian dynamics simulation methods, the fundamental mechanisms involved in facilitated diffusion of DNA-binding proteins near and on DNA chains, and the relative roles these mechanisms play in determining the mean time proteins require to find their specific targets on a DNA chain. Two different scenarios were investigated. In the first scenario, the DNA was an extended chain aligned on a surface with a cognate site located at one end of the DNA, exposed to a homogenous solution of proteins with attractive interactions for the DNA monomers. The dynamics of these proteins were characterized and mean time required for proteins to find their specific target DNA monomer was studied as a function of protein concentration in the bulk and the strength of the non-specific binding energy between DNA and proteins. An optimal binding energy was identified corresponding to the most efficient search process. In the second scenario, the DNA also had a coiled conformation. The effect of DNA conformation on protein transport and the mean first passage time was studied. Here, it was discovered that proteins found their target on a coiled DNA much faster that on partially extended DNA chains. The occurrence of inter-segmental transfers, where proteins moved a large distance in sequence space by short hops across loops, was confirmed and correlated with the enhanced target search efficiency observed in coiled DNA. The results of this simulation study reproduced some of the previous predictions of kinetic

models and experimental observations, and extended the knowledge about the target search process of DNA-binding proteins to aspects not easily studied using available theoretical and experimental methods.

Acknowledgement

My sincere appreciation goes to my academic supervisor, Dr. James Harden for his great knowledge, support and patience during my research.

I would also extend my grateful thanks to Dr. Ralf Metzler for offering me the chance to study my Master program and getting me started in the right direction.

Finally, I dedicate my thesis to my dear brother Reza and my parents to be always there for me at any worst time.

Chapter 1

Introduction

1.1 DNA

Deoxyribonucleic acid (DNA) is a molecule that contains the genetic instructions used in the development and functioning of all known living organisms[1]. DNA usually exists as a double-stranded structure, with both strands coiled together to form the characteristic double-helix. The main role of DNA molecules is the long-term storage of information needed to construct other components of cells, such as proteins and RNA molecules, and to make copies of itself (replication) [2]. Replication is a critical role of DNA in that when a cell divides, each new cell should have an exact copy of the DNA present in the old cell. The DNA segments that carry this information are called genes (specific sites), but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information. [3].

Special proteins are key participants in DNA transcription and regulation processes. For instance, Helicase is an enzyme that binds to a specific sequence on a DNA helix and initiates the unzipping of double stranded DNA into single stranded DNA [4]. This process is the first step of replication of DNA. Transcription factors or DNA polymerases are examples of proteins that bind to a specific sequence on the DNA and thereby play a role in the transcription process [5, 6]. Although DNA has two anti-parallel strands only one of the strands is used as a template during the transcription stage. During DNA

synthesis the new strand may include some errors; in order to repair mismatched sequences on the DNA, some proteins are able to distinguish the template strand from the synthesized strand. After the recognition of the error site, they scan the DNA strand and replace the mismatch sequence with the correct sequences[7].

These examples show that while different proteins have particular interactions with DNA in the cell, their functions depend on the ability of the proteins to recognize specific sequences on DNA to start a biological process. Therefore, it is very important to understand how specific segments are targeted by specific-binding proteins. At the onset, proteins are diffusing randomly in the cell environment in their search for the specific binding sequence on DNA to start their special duties. In Eukaryote cells, the DNA is packed in the nucleus, however in Prokaryote cells the DNA is totally available in solution for proteins. Therefore one might expect ordinary 3-D diffusion to play a key role in the target search process in Prokaryote cells. Three-dimensional diffusion coefficients of proteins can be estimated from Smoluchowski's equation [8], which assumes every collision is productive and proteins have to undergo many separate collisions with the DNA before finding the target. In 1970, Riggs et al. [9] performed an experiment where the **Escherichia coli** *lac* **-repressor** protein binds to $\lambda - \mathrm{DNA}$ in vitro. Their studies found that this protein discovers its specific target site on the DNA with a reaction rate of approximately $\nu = 7 \times 10^{10}$ $M^{-1}s^{-1}$. A Smoluchowski analysis for diffusive transport in the lac-repressor system gives an upper bound on association rate and the diffusion coefficient of $\nu = 10^{8}\ M^{-1}s^{-1}$ and $D = 5 \times 10^{-7}\ \ cm^{-2}s^{-1}$, respectively. The Smoluchowski estimate for the maximum association rate for a diffusion-controlled reaction is more than an order of magnitude smaller than the observed rates reported in Ref.[9]. Riggs et al. proposed that the extremely fast reaction rate was due to electrostatic binding between the positively charged DNA and negatively charged lac-repressor [9], an effect not included in the passive diffusion analysis.

Later additional mechanisms for DNA-protein interactions were proposed by Berg, Von Hippel and co-workers [10, 11]. In particular, they hypothesized that proteins are most

likely to initially bind to random DNA sites via non-specific interactions, and then execute a diffusive search for the specific target by sliding along the DNA. This mechanism requires that the activation barrier for translocation of the protein along the DNA in the nonspecifically bound state to be sufficiently small compared with the thermal energy scale. It is possible that non-specific screened electrostatic interactions between DNA and proteins could satisfy this criterion. This proposed process, which reduces the regions of the cell volume needed to be explored by diffusing proteins via three-dimensional diffusion, is called a facilitated diffusion process. Berg, Von Hippel, and co-workers [10] further hypothesized three fundamental search mechanisms are involved in the targeting process (see Figure 1.1):

1. **Three-dimensional diffusion**, in which the protein detaches from the DNA site, explores the fluid around the DNA, and later rebinds at another site near to (local hopping) or far from where it last detached;

2. **One-dimensional diffusion (sliding) along the DNA**, in which the protein can randomly transfer to adjacent sites without dissociation;

3. **Inter-segment transfer or direct transfer between DNA segments**, in which proteins execute micro- or macro-hops between two DNA sites that are quite far apart along the DNA contour but are transiently in close spatial proximity due to fluctuations of the DNA conformation.

1.2 Single molecule experiments

To observe the facilitated diffusion of proteins searching for their specific targets on the DNA, several experimental studies have been performed. However, since such measurements are extremely difficult in a real cell environment, Experiments have focused on single molecule measurements in controlled *in vitro* settings. Single molecule measurements are

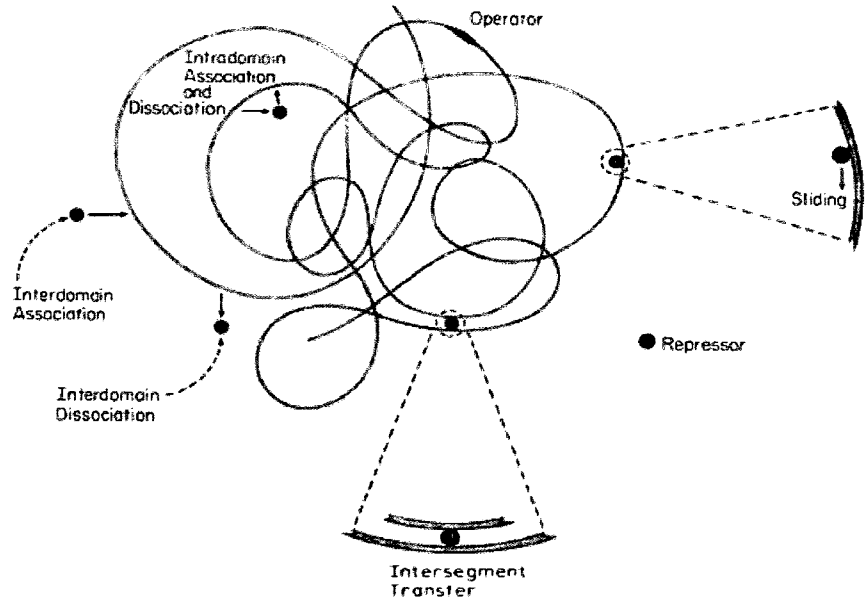


Figure 1.1: Fundamental mechanisms of facilitated diffuion of protein along DNA [10].

a class of experiments in which individual molecules (e.g. DNA and DNA-binding proteins) are isolated and individually manipulated *in vitro*. In such situations, it is possible to study and measure several different properties of isolated, interacting macromolecules that cannot be observed otherwise. One of the most important techniques used in single molecule experiments to manipulate individual macromolecules like DNA is the optical tweezers method [12, 13], alone or in combination with microfluidic environments, which are briefly discussed in the following.

1.2.1 Optical tweezers

Laser optical tweezers are instruments that use focussed laser light to trap small particles in a medium of higher dielectric constant (e.g. water). Such optical tweezers are used to manipulate and measure tension forces of macromolecules that have been tethered to colloidal particle "handles" [14]. Figure 1.2 shows a schematic of the optical trapping mechanism. Light exerts a force on objects that reflect or refract the light. However, in macroscopic objects this force is so much smaller than other forces acting on it that its effect

is not noticeable [15]. Since the objects manipulated by optical tweezers are microscopic

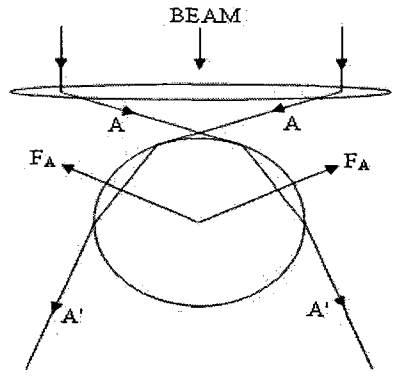


Figure 1.2: Physics of optical tweezers.

particles, they can be manipulated by the light from a very modest laser source. The object that usually is trapped by laser beam in single molecule experiments are the polystyrene or silica spheres with a size in the range of 1-3 μm. The momentum applied to such a polystyrene bead is proportional to the difference between the momentum flux entering the object and that leaving the object [16]. The single molecule experiments to stretch the DNA is carried out inside an aqueous fluid chamber consisting of two glass surfaces separated by a narrow gap, typically of order 1 μm wide. The DNA to be studied is attached at each end to a colloidal particle. Typically, one particle is immobilized on a surface or by a micropipette, while the other is captured by the optical tweezer. The DNA chain may then be elongated by moving the position of the optical trap with respect to the immobilized DNA end, as shown in Figure 1.3. Fluorescently-labeled DNA-binding proteins are then introduced into the chamber and their interaction with the DNA chain is monitored optically using a microscope, a fast digital camera, and a digital frame grabber board.

1.2.2 Shear Flow

As an alternative to using optical tweezers, another common method for manipulation of DNA chains is the use of microfluidic flow fields to stretch and align DNA in a sample chamber [14, 17]. Such experiments usually involve a simple shear flow between two

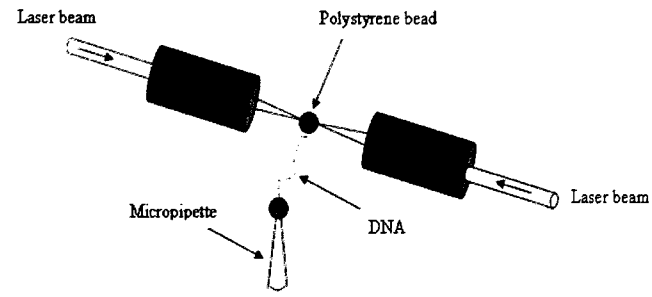


Figure 1.3: Optical tweezers used in single molecule experiments.

closely-spaced parallel plates, as shown schematically in Figure 1.4. In such a parallel plate geometry, with an upper plate moving with respect to a fixed lower plate, the fluid velocity is parallel to the plates and is a linear function of the distance y from the lower plate, as shown in Figure 1.5. DNA chains tethered by one end to the lower surface experience a drag force due to this shear flow that tends to elongate the chains in the flow direction [18, 19, 20]. The extent of elongation is governed by a dimensionless quantity, the Weissenberg number W_i, given in this case by $W_i = \dot{\gamma}\tau$, where τ is the longest relaxation time of the polymer in the fluid and $\dot{\gamma} = \frac{\partial u}{\partial y}$ (Figure 1.5) is the fluid shear rate [20, 21]. The force f applied to the polymer in the $\hat{x}$-direction scales with W_i as $f \sim W_i^{2/3}$ for the case of for the worm-like DNA chains. A sufficiently strong shear flow is required to overcome the inherent entropic preference for a coiled chain conformation and obtain a fully extended steady-state chain conformation [20].

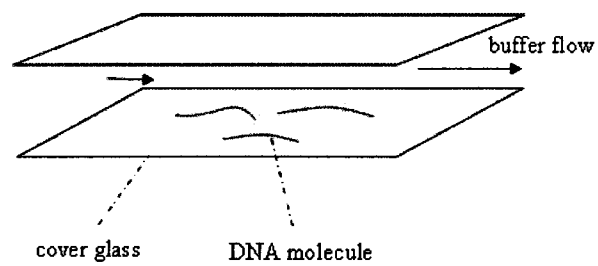


Figure 1.4: Shear flow is used to align DNA molecule on a surface.

Typically, one finds that fully extended chains can be prepared in flows with $W_i >$

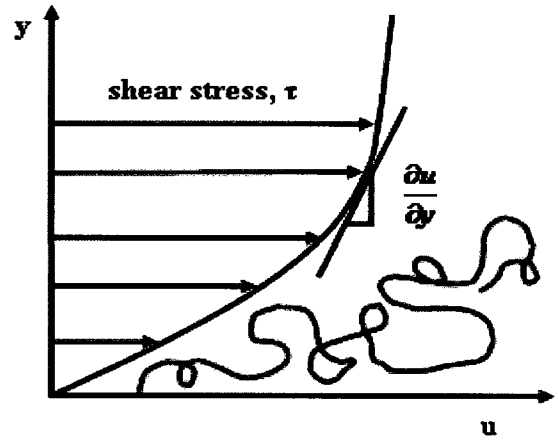


Figure 1.5: Stretching a DNA coil by shear flow.

$\mathcal{O}(10)$. In experiments on DNA target searches, fluorescently-labeled DNA-binding proteins are circulated through the flow cell at fixed concentration, where their interaction with the extended DNA chains is monitored optically.

1.2.3 Total internal reflection fluorescence microscopy (TIRFM)

Single molecule studies of target search by DNA-binding proteins require methods to precisely visualize single molecules (proteins and DNA) and their relative positions as a function of time. For the case of DNA chains localized near a surface (e.g. by laser tweezers or shear flow), Total Internal Reflection Florescent Microscopy (**TIRFM**) has proved to be one such useful method.

In TIRFM, a laser beam is guided from below through a glass cover surface at an angle greater than the critical angle for total internal reflection (see e.g. Figure 1.6). Although the light is totally reflected at this angle, the reflected beam generates an evanescent electromagnetic field adjacent to the interface. This evanescent wave decays exponentially into the optically less dense medium (e.g., water) with a characteristic penetration depth of 100 nm from the interface [22, 23]. Because the evanescent field is restricted to the region of the

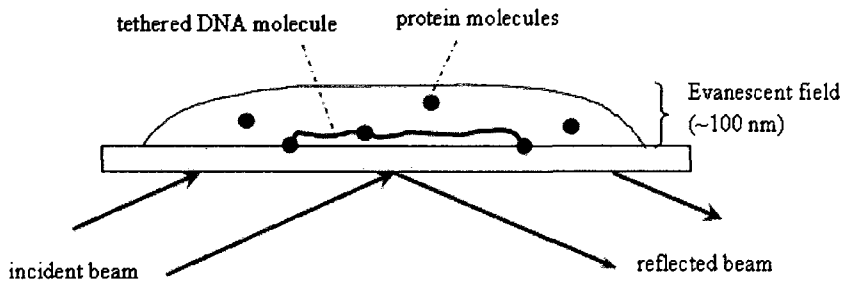


Figure 1.6: Total internal reflection fluorescence microscopy set-up.

interface, fluorescently labeled molecules (DNA and proteins in this case) will be excited only if they reside in this region. This provides a natural method for monitoring protein diffusion in the neighborhood of DNA chains that are confined to the chamber surface, as sketched in Figure 1.6.

1.3 Recent single molecule experiments

Several recent single molecule experiments using total internal florescence microscopy to visualize and methods to stretch DNA (optical tweezers and shear flow) have been employed to test the prospective mechanisms impacting the target search process. To date, direct confirmation of the Von-Hippel facilitated diffusion hypothesis has proved elusive, presumably due in part to the difficulty of direct visualization of inter-segmental transfer events in single molecule experiments. One such attempt investigated the interaction of an enzyme (EcoRI) with nine linear fragments that varied in the number of base pairs between 34 and 6200 [24]. If the key inter-segmental transfer events associated with facilitated diffusion were essential, the binding rate would not be a monotonic function of the length of the DNA fragments. However, this experiment indicated that the EcoRI target discovery time was inversely correlated with the length of the DNA fragment, suggesting a pure sliding mode of target search. In the following, we briefly outline some of the other key single molecule studies and their findings.

1.3.1 Hopping versus Sliding

Many single molecule experiments have been conducted that were able to distinguish between sliding and hopping modes of protein motion on and around DNA substrates. Here we list a few examples:

(1) One experiment visualized the diffusion of Human oxoguanine DNA glycosylase 1 (hOgg1) on double stranded λ-DNA stretched to an extended state 16 μm in length via shear flow [25]. The net displacement of individual molecules indicated, that despite the buffer flow, the diffusion was not biased in the flow direction. Using proteins that do not consume biochemical energy for their diffusion, this study found that the target search along the DNA was due to thermal motion and was a directionally unbiased process. (We note that some other studies reported convective motion of DNA binding proteins [26, 27], presumably the result of excessive drag forces due to solvent flow dominating over thermal forces in these experiments). A linear relationship between the mean square displacement versus time was verified, with a one-dimensional diffusion coefficient $D_{1d} = 4.8 \times 10^6$ bp^2/s, a value consistent with sliding motion that also involves protein rotation around the DNA helix. The data also predicted an activation barrier for translocation along DNA to be about 1 k_BT, consistent with theoretical predictions [28]. Increasing salt concentration in this system had the effect of increasing the non-specific binding affinity of hOgg1 for DNA. Interestingly, such an increase in salt concentration did not affect the measured target search time, strongly suggesting that sliding rather than hopping dominated the process.

(2) Gorman et al. used TIRFM to visualize the diffusion of Mismatch repair complex Msh2-Msh6 on λ-DNA [29]. This experiment relied on biotinylated λ-DNA (with 48,502 bp) that was tethered by both ends on a sample chamber surface otherwise coated with a lipid bilayer (see Figure 1.6 above). Inter-segmental transfer was unlikely in this case because the DNA was maintained in a physically extended configuration. Hoping and sliding could be distinguished from one another by analyzing the change in the behavior of the effective diffusion coefficient with different concentrations of salt. The life time of the bound Msh2-Msh6 decreased with increasing concentration of salt, but there were no significant changes

observed in the mean diffusion coefficients. This result suggests that proteins are very unlikely to experience large hops as they travel on the DNA. The data from 125 protein trajectories revealed a broad distribution of diffusion coefficients with a mean value of 1.2×10^{-2} μm^2/s and a range of values from 2×10^{-4} μm^2/s to 9×10^{-2} μm^2/s. Such a broad range may indicate that there are transiently trapped proteins.

(3) In several studies, hopping motion was found to be a dominant process. For instance, Halford et al. used DNA substrates with two cleavage sites for the BbvCI enzyme separated by different sequence lengths and with varying orientations relative to one another [30]. Such an approach is designed to reveal whether the protein maintains contact with the DNA as it transfers from one site to another by sliding or whether it loses and subsequently regains contact by a dissociation/re-association step. In the study of Ref. [30], two different orientations of DNA molecules were examined: (i) repeated sites and (ii) inverted sites. If the translocation of an enzyme from one site to another happens only by sliding, then the enzyme could cleave the DNA in both repeated sites, but it should not cleave both sites if the sites were inverted. However, if the translocation involves a dissociation step, the probability of DNA cleavage on repeated sites will equal that on inverted sites. Results for reactions at low salt concentration indicated that the test protein stayed on the DNA as it traveled between sites provided the sites were <50 bp apart. On the other hand, transfers of >30 bp in physiological salt conditions, or over distances of >50 bp at any salt concentration, always included dissociation steps. Hence, for this enzyme 3-D detachment/attachment is its main mode of translocation.

(4) Another selective cleavage experiment utilized an enzyme that can bind and cleave at two distinct recognition sites on double stranded DNA [31]. In this experiment, these two specific recognition sites were separated by n base pairs in the middle of a double stranded DNA fragment. If the primary transport mechanism of the enzyme is hopping via three-dimensional diffusion, then binding and cleaving at the two sites would be independent processes occurring with roughly equal probabilities. On the other hand, If sliding via 1-D diffusion was primary transport mechanism of the enzyme, cleavage at the two binding

sites would be a correlated process. The two mechanisms would yield different fragment distributions. These were characterized by a processivity factor that was defined as the number of the reactions in which the enzyme cleaves both sites before departing from the domain of the DNA relative to the total number of reactions. The experiment with the same enzyme was tested on 4 pairs of DNA substrates containing two sites separated by various distances. The processivity factor was found to decline as the length n of DNA between the sites was increased, indicating that hopping was the dominant mechanism for this system.

1.3.2 Movement of protein along DNA chains

A number of measurements have focused on characterizing the nature of the motion of proteins along the DNA. The local details of this motion can strongly impact measured transport properties. For some very tightly bound proteins, sliding diffusion involves rotation around the DNA chain axis, in order to remain in a major or minor groove. Other proteins that are less strongly bound avoid this constraint by making short micro-hops. Theoretical studies [8] have suggested that for a typical protein size there may be up to a 1000 fold difference between the effective diffusion coefficients for such rotating and hopping modes transport. Several single molecules experiments conducted to obtain 1-D diffusion coefficients for proteins moving along DNA chains are described below:

(1) One-dimensional diffusion of the T7 RNA polymerase along stretched DNA molecules was visualized in a single molecule experiment using TIRFM on aligned DNA [32]. The DNA combing method, in which a polystyrene-coated cover glass was dipped into a DNA solution for a short period of time and then pulled out at a constant rate, was used to stretch and anchor the DNA on a substrate. A protein initially bound to the DNA diffused freely for as long as several minutes, providing sufficient data to observe diffusive motion of the proteins. The resulting 1-D diffusion coefficient of proteins varied from molecule to molecule over a large range of values from 6.1×10^{-11} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ to 4.3×10^{-9} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$, values consistent with rotational diffusion coefficients estimated by theory [8].

(2) Another single molecule experiment on mismatched proteins visualized the diffusion of human Rad51 on double-stranded DNA [27]. The double-stranded DNA molecule was tethered to the lipid bilayer surface by one end, extended along the surface using hydrodynamic drag, and finally attached by the distal end to the surface. The observed motion of Rad1 on the extended DNA in the absence of flow was unbiased, bidirectional 1-D diffusion, with the proteins moving freely on the DNA for a long period of time. The mean square displacement of the proteins was found to be a linear function of time interval, as expected for 1-D diffusion. The inferred average 1-D diffusion coefficient was 0.042 ± 0.054 $\mu m^2/s$ during a period of 124 seconds, in which the sliding length ranged from 50-150 μm. (3) A recent single molecule experiment by Wang et al. confirmed that Lac repressor could locate its target via the sliding mechanism and provided a direct experimental explanation for the observation of Riggs, et al. discussed previously. In this work, 1-D Brownian motion of *LacI* repressor proteins on a DNA contour was visualized on a single molecule using the TIRFM method[33]. Hydrodynamic flow was used to stretch the DNA to 90% of its total contour length. The diffusive motion of subsequently bound *GFP-LacI* covered a length of the DNA contour from 120 nm to 2920 nm, leading to a range of calculated 1-D diffusion coefficients from 2.3×10^{-12} cm^2/s to 1.3×10^{-9} cm^2/s. In this experiment, there was an apparent lack of correlation between one-dimensional diffusion coefficients and sliding lengths, which could have been due to conformational distributions in the protein, but which could not have been directly visualized in the experiment.

1.3.3 Inter-segmental transfer

One of the proposed facilitated diffusion mechanisms is inter-segmental transfer or direct transfer of proteins between DNA segments that are far apart in sequence space, but transiently brought into close spatial proximity by a fluctuation of the DNA conformation [11], as sketched below in Figure 1.7. This process was not dominant in the previous experiments reviewed because extended DNA coils do not have distant parts of the sequence in close proximity. Intersegmental transfer processes are most likely to occur for coiled DNA

and proteins with specific affinity for two or more target sites in the DNA sequence, such as the lac repressor, human Hox-D9 homeodomain, and gluocorticoid receptor proteins.

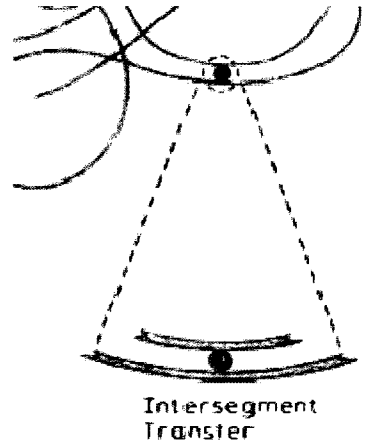


Figure 1.7: Intersegmental transfer of protein on DNA [10].

The existence of intersegmental transfer has been confirmed by several experiments. The experiment of Iwahara et al. [34] used human Hox-D9 homo-domain proteins that have two distinct DNA binding sites and that can transfer directly between DNA segments. The NMR Exchange spectroscopy method was used to detect inter-segmental transfer events. The study reported indirect yet unambiguous evidence for protein transfer events between distinct binding domains in solutions of DNA and enzyme. A recent single molecule experiment has studied the impact of DNA conformation on the specific protein-DNA association rate of proteins thought to be involved in inter-segmental transfers [35]. In this study, the conformation of a 6500bp length DNA with a central target recognition site for the EcoRv enzyme and colloidal particles attached to the termini was manipulated by optical tweezers. The association rate of EcoRv enzyme on single DNA molecules, as determined by DNA cleavage events, was computed from 385 cleavage reactions as a function of DNA extension. Three levels of extension (straight, random coil and a squeezed coil) were created by changing the distance between optical tweezers beads at the DNA chain termini. The maximal association rate was found in a random coil configuration, while the association rates

for both extended and squeezed DNA were found to be very low. The difference between association rates for the different DNA configurations were related to the suppression of inter-segmental jumping in extended DNA [35, 36]. Since non-specific binding of proteins to DNA is mostly due to electrostatic interactions, the concentration of salt was expected to modify the association rates at different DNA extensions. At very low and very high salt concentrations, there was no significant difference in the association rates for either coiled or extended DNA, but an intermediate salt concentration ($\sim$ 60 mM) yielded an optimal difference between the association rates for coiled vs extended DNA, an effect attributed to favorable conditions for inter-segmental transfer.

1.4 Theory of facilitated diffusion

As experimental single molecule experiments of the facilitated diffusion progressed, other research groups have used different theoretical and computational methods to investigate the facilitated diffusion process. While limited in scope and molecular detail, such theoretical methods can provide insight into the various potential mechanisms involved in this process. Below a few key theoretical studies are summarized.

1.4.1 Target recognition on a stretched DNA chain

A number of computational and theoretical studies have been made of the facilitated diffusion of proteins on stretched DNA. These studies treated the proteins and DNA subunits as structureless or coarse-grained entities.

(1) In a study of protein diffusion in 1-D, Sokolov et al. [37] considered the dependence of the target search process on the concentration of proteins. Whereas in pure three-dimensional diffusion the specific binding rate is directly proportional to the concentration of proteins in a solution (C), this work predicted that the rate of specific binding was proportional to C^2 for a pure one-dimensional diffusion process with excluded volume. A scaling method was employed to explain the observed dependence of the mean first passage

time (MFPT) on C. The MFPT is the mean time needed for a protein to initially find its target sequence on a DNA molecule. In this study N particles were initially placed randomly on a straight line of length L and allowed to fluctuate in position under the constraint that two particles may not occupy the same position. The target was located at one end of the line and the MFPT to reach this target was obtained as a function of N in studies of 10^5 single events. The MFPT was found to scale with protein concentration $C \sim N/L$ as MFPT$\sim C^{-2}$, in agreement with one of the single molecule experiments [38]. However, in realistic experiments it is not possible to have pure-one dimensional diffusion, since thermal fluctuations will always generate protein hops off the DNA chain. However, biochemical conditions can often be imposed that minimize the rate of unbinding, leading to quasi 1-D diffusion of proteins occurring within a thin shell around the DNA chain.

(2) Computational tools were used to explore the facilitated diffusion of three helical binding proteins (Engrailed Home domain HoxD9, Sap1 and Skn1) and one non-DNA binding protein (SH3) along stretched DNA [39]. In this method, a 100bp-B DNA was modeled by three beads per nucleotide and the proteins were represented by one bead for each amino acid residue. Non-specific protein-DNA interactions were modeled by electrostatic interactions between negatively charged DNA phosphate beads and the positive charges of protein residues. Inter-segmental transfer could not be studied using this model because of the straight conformation of the DNA chain. The trajectories of both binding and non-binding proteins were studied at low salt concentration. In these conditions, the binding proteins remained close to the DNA main axis, indicating that translocation was mainly one dimensional. In contrast, the non-binding SH3 showed no attraction to DNA and therefore executed random 3-D diffusive motion around it. The fractions of sliding, hopping and three-dimensional diffusion during the specific target search was studied at different salt concentrations for two different temperatures, $T = 0.5T_f$ and $T = 0.9T_f$, where T_f is the protein's folding temperature. As salt concentration increased, the affinity of the binding proteins for the model DNA decreased, thereby decreasing the fraction of sliding and increasing the fraction of 3-D diffusion. Interestingly, the fraction of hopping exhibited

a non-monotonic dependence on salt concentration, with a sharp peak at about 0.06 M at $T = 0.5T_f$ and 0.04 M at $T = 0.9T_f$. This verified the expectation that an intermediate binding strength is required for hopping transport to occur. As a conclusion, these authors found that optimal search efficiency was achieved when sliding constitutes 20% of the total search process.

1.4.2 MFPT and non-specific binding energy

Previously discussed experimental studies clearly showed that non-specific binding plays an important role in controlling the balance between one-dimensional diffusion and three-dimensional diffusion processes. Numerous theoretical studies have investigated the dependence of the search process on the non-specific binding energy and the interplay between 1-D and 3-D diffusive processes. These studies typically involve continuum kinetics calculations that treat the dimensionality of the problem (a 1-D curvilinear chain contour embedded in 3-D) in an implicit fashion. two such models are reviewed below:
(1). A kinetics-based approach was used by Slutsky and Mirny to study the facilitated diffusion of proteins on a DNA contour with different sequence-dependent potentials [28]. This model estimated the optimal time of a search process, including a combination of both one-dimensional and three-dimensional diffusion. The calculated optimal time corresponded to trajectories for which a protein spent approximately half of its time sliding along the DNA and the other half diffusing in three dimensions. It was also revealed that the non-specific binding energy plays an important role in controlling the balance between the two processes. For a sufficiently rough free energy landscape, diffusion proceeds too slowly to efficiently scan the available binding sites on the DNA in a realistic period of time. In contrast, if the landscape is very smooth, only three-dimensional diffusion would occur, pre-empting the facilitated diffusion mechanism. An optimal non-specific binding energy of $\epsilon \simeq 2k_BT$ was suggested by this work, representing an optimally rough free energy landscape for facilitated diffusion. On the other hand it was claimed that after a protein finds its target, it needs to have enough time for a change in the protein conformation to

achieve its function, which requires quite stable DNA-protein binding. Paradoxically, realistic energy functions cannot provide both rapid search at intermediate binding strength and strong binding of a protein. Thus, it was suggested that two modes of protein-DNA binding co-exist: the search mode and the recognition mode. A protein in the non-specific binding search mode is not aware of the details of the DNA sequence it is transiently bound to. Therefore, at each site it must switch to a specific binding recognition mode that probes the visited sites for its level of specificity and undergoes a change in conformation when the target site is discovered.

(2) Another kinetics-based approach argued that the target search process of proteins on DNA involves three mechanisms: three-dimensional diffusive motion, one-dimensional diffusive motion, and the correlation of the two [40]. The goal of this work was to develop a qualitative picture of the search and detection of targets on DNA. A key result of the analysis was an expression for the search time τ_c as a function of these three processes:

$$\tau_c = \frac{x^2}{2D_3} + \frac{\lambda^2}{2D_1} + \frac{x\lambda}{D_1 y_{eff}} \tag{1.1}$$

The first two terms correspond to the time spent by the protein in 3-D and 1-D diffusion, with diffusion constants D_3 and D_1, respectively. The last term is a correlation term accounting for the contributions of trajectories for which a protein goes from a 3-D to a 1-D transport state but unbinds from the DNA before it travels a full sliding length λ. The two characteristic distances, x and λ, are the average distance of a protein in solution from the DNA chain (related to the concentration of proteins) and the average sliding length along the chain, respectively. The dimensionless parameter y_{eff} is a measure of the non-specific binding energy between proteins and DNA. The last term in Eq.1.1 partially accounts for fluctuations in the length of the 3-D and 1-D trajectories. Further analysis uncovered a reduction in the search time achieved at some intermediate strength of the protein-DNA binding energy and concentration of free proteins. If the concentration of free proteins in solution was very low, the correlation term dominates and the relative search time becomes

very long. This means that the protein molecule spends a small fraction of its time on the DNA, frequently binding and rebinding during the scanning process. In the opposite limit, where the concentration of protein is large, there is no need to slide along the DNA, as there is always a protein near enough to find the target via 3-D diffusion.

1.4.3 Mean first passage time and DNA conformation

A number of theoretical studies have derived estimates for the mean first passage time of proteins searching for a target on a DNA coil. Generally, these studies have considered a combination of at least two different searching mechanisms, 1-D diffusion and 3-D diffusion. Two examples are briefly discussed below.

(1) Halford et al. [30] derived a phenomenological kinetics equation providing a quantitative estimate for the reaction time of a protein that is moving under the control of two competitive transport mechanisms in a crowded environment. The resulting reaction rate k per unit protein concentration $C = N/V$ had the form:

$$k = (\frac{1}{Dl_{sl}} + \frac{Ll_{sl}}{D_1 V})^{-1} = Da(\frac{a}{l_{sl}} + \frac{D}{D_1} aLl_{sl}C)^{-1} \tag{1.2}$$

where N is the total number of proteins per DNA, V is the total volume of the DNA coil, a is a local molecular volume, L is the contour length of the DNA, l_{sl} is the sliding length, D is the effective 3-D diffusion constant, and D_1 is the effective 1-D diffusion constant. The first term accounts for the 3-D diffusion-limited rate (Da) and the final factor represents the acceleration of the reaction. As sliding length increases, this expression predicts that the 3-D diffusion term decreases and the total time spent executing 1-D sliding diffusion increases. By balancing the two terms in Eq. 1.2 with respect to sliding length, one obtains a maximum reaction rate at an optimal value of sliding length:

$$l^*_{sl} = \sqrt{\frac{D_1 V}{DL}} \ \ and \ \ \tau^* = 2\sqrt{\frac{LV}{DD_1}} \tag{1.3}$$

(2) Kleinin et al. [41] presented a complementary quantitative model, also based on consideration of both 3-D diffusion and 1-D sliding along DNA. This analytical model provided predictions for the mean first passage time, verified by numerical simulations. Their prediction for the reaction time resulting from the combined effects of 1-D and 3-D diffusion is given by:

$$\tau = \frac{V}{8D_{3D}\xi} + \frac{\pi L \xi}{4D_{1D}} \tag{1.4}$$

where V is the confinement volume, L is the DNA chain length, D_{1D} and D_{3D} are the 1-D and 3-D diffusion constants, respectively and ξ is a characteristic distance where 1-D and 3-D diffusive processes are in balance. The non-monotonic dependence of τ on ξ confirms that non-specific binding of protein to DNA can reduce the reaction times significantly. In particular, a minimum search time τ_0 can be found at an optimal value of ξ for a variety of conditions. Note that Eq. 1.4 is very similar to an expression reported by Halford and Marko [30], if ξ is identified with the sliding length l_{sl}.

1.5 Thesis Overview

This thesis presents a simulation study of the target search process described in the experimental, computational, and theoretical studies reviewed above. Unlike some of the previous theoretical and computational studies, this work explicitly considers the behavior of diffusing protein molecules interacting with fluctuating DNA chains in three dimensions. Chapter 2 describes the coarse-grained model representing the DNA chain, the protein molecules and the Brownian dynamics method used to simulate the system. Chapter 3 presents simulation studies of stretched DNA chains interacting with diffusing proteins, and characterizes the 1-D and 3-D aspects of protein transport and their effect on the target search process for extended DNA chains. Finally, Chapter 4 presents comparative simulation studies of

stretched and coiled DNA chains, with a focus on the effects of inter-segmental transfers on the protein dynamics and the target search process.

Chapter 2

Model and Simulation Methods

In the following chapter, we describe the methods used in the simulation studies of facilitated diffusion of proteins along a DNA chain. We first briefly describe the LAMMPS simulation package and the Brownian dynamics method used to simulate the behavior of a model interacting DNA-protein system. We then describe the coarse-grained model of the DNA chain as a bead-spring polymer and the proteins as coarse-grained single beads or aggregates of beads, and give some details about the simulation protocols used for protein searches of specific targets on DNA.

2.1 Brownian Dynamics Simulation

Brownian Dynamics (BD) simulation is a mesoscopic simulation method for molecular species in solvent which avoids treating the solvent molecules explicitly by implicitly incorporating their effects on solute particles in the equation of motion of particles [42, 43, 44]. In a system with a relatively low concentration of solutes, this method allows for simulation of the slow dynamics of solutes over much longer time scales than in a simulation of all molecules by replacing the the effects of rapid collisions on the solvent molecules by random thermal forces and an associated viscous damping force. In essence, BD corresponds to the standard Langevin dynamics method in the overdamped limit where no average ac-

celeration takes place during the simulation and the total force on each solute molecule is always approximately zero.

2.1.1 Interparticle Interactions

Total force acting on a particle in a BD simulation is composed of four different forces: $\vec{F}_{tot} = \vec{F}_c + \vec{F}_f + \vec{F}_r + \vec{F}_{ex}$. In the overdamped limit, we take $\vec{F}_{tot} = 0$ as a given. $\vec{F}_{ex}$ is the total external force acting on the particle. $\vec{F}_c = -\vec{\nabla}U$ is the conservative force computed from the total inter-particle interaction potential, in principle a function of all particle coordinates $\{\vec{r}_1, \vec{r}_2, ...\vec{r}_N\}$ (to be discussed below). In addition to these standard forces encountered in any classical molecular simulation, there are two additional interrelated forces that are particular to BD. $\vec{F}_f$=-$\gamma\vec{v}$,is a viscous damping force which is proportional to the instantaneous solute particle velocity, where γ is a phenomenological damping coefficient related to the thermal energy scale k_BT and the solute self diffusion coefficient D via the Einstein relation: γ=k_BT/D. F_r is a stochastic force due to collisions of the solvent molecules with the solute particles. This stochastic force is taken from a random distribution satisfying the fluctuation-dissipation theorem: $\langle\vec{F}_r(t)\rangle = 0$ and $\langle\vec{F}_r(t)\vec{F}_r(t')\rangle = 2k_BT\gamma\delta(t-t')\mathbf{I}$, where $\mathbf{I}$ is the unit tensor. This leads to forces applied in random directions of RMS magnitude F_r=$\sqrt{2k_BT\gamma/\Delta t}$, where Δt is a characteristic microscopic collision time, which we take as a fundamental time-step for the simulation. The damping coefficient (γ) is always positive, thus the temperature of the system will decrease during the simulation due to the viscous damping force. This is balanced by the random force term, which is controlled in the simulation using a Langevin thermostat.

2.1.2 Dynamics Algorithm

Brownian dynamics simulation is implemented using the LAMMPS package [45] in an NVT ensemble. The NVT ensemble is a canonical ensemble in which the number of particles, the volume of the simulation region and the temperature is fixed to specified values. LAMMPS is a classical molecular dynamics code developed at Sandia National Labs in

the United States that is optimized for use on massively parallel computer clusters using the MPI parallelization scheme. It is distributed as open source code under the terms of the GPL and can be freely downloaded from lammps.sandia.gov. LAMMPS is a flexible package compatible with a variety of force fields for classical simulations at the atomic, molecular, and coarse-grained mesoscopic scales. We use the standard LAMMPS implementation of BD without hydrodynamic interactions between particles. In this case, the position of each particle i is updated using an explicit forward-time Euler scheme:

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \vec{v}_i(t)\Delta t \tag{2.1}$$

where $\vec{v}_i(t)$, the velocity of particle i, is determined at time t by global force balance using the known positions of all particles $\{\vec{r}_1(t), \vec{r}_2(t), ...\vec{r}_N\}$ at time t via:

$$\vec{v}_i(t) = \frac{1}{\gamma}\left[-\vec{\nabla}_i U(\{\vec{r}_1(t), \vec{r}_2(t), ...\vec{r}_N(t)\}) + \vec{F}_{ex}(\vec{r}_i(t)) + \vec{F}_r(\vec{r}_i(t))\right] \tag{2.2}$$

To insure stability of the explicit time integration scheme, we use a small timestep of $\Delta t = 0.005\tau$ where τ is the natural time unit of the BD simulation.

2.1.3 LJ reduced units

Typically, physical quantities measured in BD simulations are represented by dimensionless, or reduced, units defined by the characteristic particle size σ, particle mass m, and the simulation energy scale ϵ (determined by the interparticle potential). Using these characteristic values, we may obtain dimensionless quantities for time ($\sqrt{m\sigma^2/\epsilon}$), pressure ($\epsilon/\sigma^3$), and density ($m/\sigma^3$). From these characteristic quantities, we define non-dimensional system variable constructed from:

$$\begin{aligned}
\text{dimensionless length:} \quad & \tilde{r} \equiv r/\sigma \\
\text{dimensionless energy:} \quad & \tilde{E} \equiv E/\epsilon \\
\text{dimensionless time:} \quad & \tilde{t} \equiv t/\sqrt{m\sigma^2/\epsilon} \\
\text{dimensionless pressure:} \quad & \tilde{p} \equiv p/(\epsilon/\sigma^3) \\
\text{dimensionless density:} \quad & \tilde{\rho} \equiv \rho/(m/\sigma^3)
\end{aligned}$$

2.1.4 Periodic Boundary Conditions

We employ periodic boundary conditions [46] throughout this study. In the periodic boundary scheme, the unit cell is replicated infinitely in one or more directions. Whenever a particle leaves the unit cell simulation region by passing though one boundary, it automatically re-enters the region through the opposite boundary, thereby creating a periodic image of itself and keeping the concentration of particles in the unit cell constant. Periodic boundary conditions are utilized in situations where we wish to avoid molecular interactions with a wall or avoid restricting the degree of translational freedom of the particles. In Chapter 3, we discuss DNA chains stretched and pinned to a wall (as in the case of a microfluidic cell). In this case, we had an upper/lower wall pair and periodic boundary conditions in the two transverse directions. In Chapter 4, DNA chains were located in the central region of the simulation cell and periodic boundary conditions were applied in all three directions.

2.2 DNA Model

The DNA chain molecule is modeled as a bead-spring polymer with N coarse-grained beads of the size σ interconnected by $N-1$ springs. The spring models the entropic restoring force associated with stretching a subsection of the chain, as indicated in Figure 2.1. In addition to bonded interactions between neighboring beads, each DNA monomer bead interacts with other non-neighbor beads (and also with the protein species described below) via the standard Lennard-Jones (LJ) potential:

$$V_{bond}(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{2.3}$$

where ϵ is the depth of the potential wall, σ is the (finite) distance at which the inter-particle potential is zero, and r is the distance between the particles. The r^{-12} term describes Pauli repulsion at short ranges due to overlapping electron orbitals, and the r^{-6} term describes attraction at long range due to van der Waals or dispersion-like forces. Lennard-Jones

potential is plotted for $\sigma = 1$ and $\epsilon = 1$ in Figure 2.2. Note that the minimum value of the LJ potential occurs at $r = 2^{\frac{1}{6}}\sigma$.

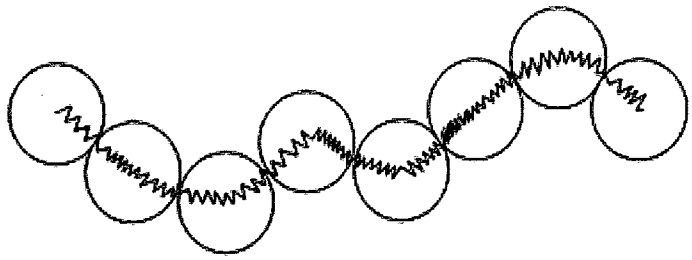


Figure 2.1: Bead-spring polymer.

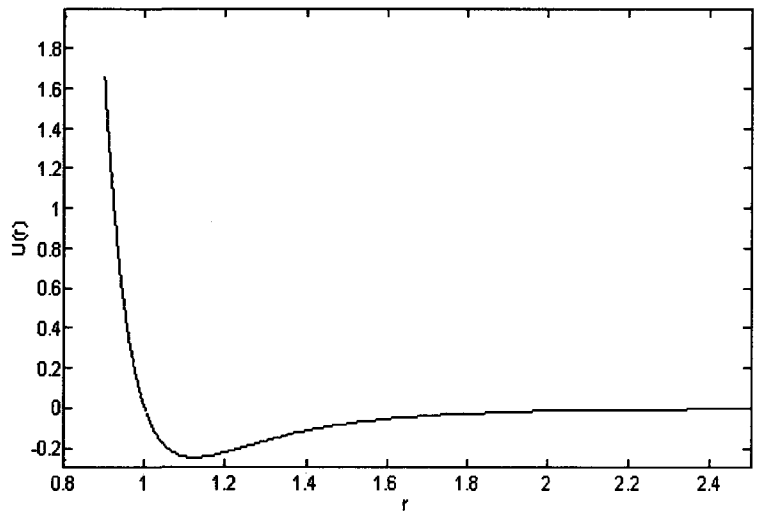


Figure 2.2: Lennard Jones (LJ) potential.

The bonding between DNA monomers is described by a hybrid potential which includes a FENE (Finitely Extended Nonlinear Elastic) bond potential plus a shifted LJ (Lennard-Jones) potential:

$$V_{bond}(r_{ij}) = -0.5\kappa R_0^2 ln\left[1 - \left(\frac{r_{ij}}{R_0}\right)^2\right] + 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] + \epsilon \qquad (2.4)$$

R_0 in the first term of the hybrid potential is the maximum extent of the Fene bond, r_{ij} is the distance between i and j monomers, and κ is the FENE spring constant. Excluded

volume interactions are introduced via a pure short-ranged repulsive LJ potential with a cutoff radius of $r = 2^{\frac{1}{6}}\sigma$ and an additive energy shift ϵ. In addition to a bonding potential, we also utilize a harmonic bending potential to modulate the intrinsic stiffness of the DNA chain. This potential has the standard form:

$$V_{angle}(\theta - \theta_0) = \kappa_b (\theta - \theta_0)^2 \tag{2.5}$$

where θ is an equilibrium value of the angle between subsequent bonds and κ_b is a bending constant. Note that the usual factor of 1/2 in the harmonic bending potential is included in κ_b. Figure 2.3 illustrates the bending associated with finite angular deformation of the DNA chain.

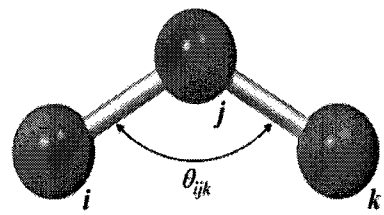


Figure 2.3: Angular deformation and bending potential.

2.3 Protein Models

Facilitated diffusion of proteins along a DNA chain is investigated via two different protein molecular models. For studies of extended DNA chains (Chapter 3) it is sufficient to use a simple spherical bead model of the protein molecules with attractive LJ interactions with the DNA monomers and repulsive LJ interactions between protein molecules. For studies of coiled DNA chains (Chapter 4), a more elaborate multi-bead protein model is required, consisting of an attractive LJ head bead partially surrounded by a shell of repulsive LJ beads. Figure 2.4 shows depictions of these two types of protein beads. Their detailed features are described separately below.

Figure 2.4: protein as a multi-bead shell (left),Protein as a single bead (right).

2.3.1 Proteins as single beads

In Chapter 3, studies of the target search process are presented for a DNA molecule that was stretched into an extended, linear conformation. For these studies, the searching proteins were considered as spherical beads of size 2σ, i.e. twice the size of DNA monomers. The protein bead size was chosen to be larger than the DNA monomers to facilitate the sliding diffusive motion of a bound protein along the DNA chain. This choice also reflects the size asymmetry for DNA binding proteins and their DNA substrates. The interaction between protein molecules was a purely repulsive LJ interaction with a cutoff radius of r_c=$2^{\frac{1}{6}} \times 2\sigma$, while the interaction between protein beads and DNA monomers was attractive with a cutoff radius at r_c= $2.5 \times 1.5\sigma$. For this choice of parameters, the average distance R between a bound protein molecule of size 2σ and a DNA monomer of size σ was $R = 1.5\sigma$.

2.3.2 Proteins as compact shells

In Chapter 4, studies of the target search process are presented for a coiled DNA molecule. In such a fluctuating random coil, distant monomers along the chain are occasionally in close proximity (e.g. when a transient loop forms). Since the LJ interaction is not a directional one, in this case one attractive LJ particle may simultaneously bind to two or more DNA monomers, leading to transient intra-molecular cross-linking of the DNA chains by the proteins. In order to favor exclusive binary interactions between protein molecules and DNA monomers and to avoid unphysical cross-linking events, a different protein model was developed. In this case, the protein molecule was chosen to be a compact shell-like

arrangement of 6 coarse-grained LJ beads, each of of size σ. One of the beads (the gray bead in the left panel of Figure 2.4) was attractive for the DNA monomers, while the others were repulsive for the DNA monomers and each other. All six beads were strongly bound to each other by a harmonic potential, $u(r_{ij})$=$0.5\kappa(r_{ij} - r_0)^2$, with a large spring constant $\kappa = 500\epsilon/\sigma^2$ and an equilibrium separation $r_0 = 0.92\sigma$, in order to minimize fluctuations in the protein conformation during the simulations.

2.4 Length and time scales

In the simulations described in this thesis, the size of the proteins (2σ) can be compared with the typical size of a protein (about 10 nm) to establish a connection between computational and experimental length scales. Specifically, each bead corresponds to about σ=5nm. For DNA , 3 base pairs is about 1nm, implying that a 5nm bead size is about 15 base pairs. A protein of size 10 nm with a three-dimensional diffusion coefficient of D_{3d}=4×10^{-7} $\frac{cm^2}{s}$ requires about 4.5×10^{-7} to move a distance of its own size. In the simulations to be discussed in this thesis, it takes about 11350 time steps for such a model protein to move a distance of 2s (10 nm). Therefore, one time step of the simulation corresponds to approximately 4×10^{-10}s.

2.5 Simulation Procedure and Analysis Programs

Running a simulation using the LAMMPS package requires several input files with information about the initial particle coordinates and the simulation parameters. Two samples of LAMMPS input scripts are given in Appendix A: one for the bead-like proteins with a stretched DNA chain, and the second for shell-like proteins with a coiled DNA chain. A new code was written to create files of initial protein and DNA bead coordinates and velocities needed to initialize the BD simulations described in the following chapters. a sample of these initialization programs are given in Appendix B. During simulations runs, particle coordinates at selected time points were written to files for subsequent analysis using

the LAMMPS "dump" command. Analysis programs were written that use such data to characterize the protein dynamics and interactions with the DNA monomers. Samples of these are included in Appendix C. One of the most important quantities obtained in these simulations was the mean first passage time (MFPT), the mean time required for the first discovery of a target DNA monomer by a diffusing protein particle. The method for calculating this quantity involved the averaging of results for a very large number of independent simulation runs (about several thousand) using the programs included in Appendix D and as described in the following chapters.

Chapter 3

Proteins and Stretched DNA

As almost all single molecule experiments have used a stretched DNA to visualize the one-dimensional diffusion of binding proteins, in this chapter we study the diffusion of proteins along a stretched DNA aligned on a two-dimensional surface. After characterization of the association of the model protein elements with the DNA chain and their motion along the DNA chain, this study focused on determination of the mean time needed for the first incidence of a protein initially in solution to find a specific target on the DNA (the so-called Mean First Passage Time) as a function of bulk solution protein concentration and the binding energy between DNA and proteins.

3.1 Model

In the spirit of a coarse-grain representation, the target DNA molecule was modeled as a bead spring polymer comprised of $N = 100$ beads of size σ(i.e. a total of 1500 base pairs) joined together by a modified FENE bond potential of the form given in Eq. 2.4 with spring constant $K = 100\sigma/\epsilon^2$. The strong spring constant restricted the inter-bead bond length to $R < R_0 = 1.22\sigma$. The repulsive part of the bond potential had cutoff radius of $r_c = 1.1224\sigma$ to prevent particle overlap.

The diffusion of protein molecules was first studied on an extended DNA chain, which was prepared by applying a force of magnitude $f = 10\epsilon/\sigma$ in the $\pm\hat{x}$-directions to the two end beads of an initially relaxed DNA coil. In order to assist the stretching process, a bending potential with spring constant $K_b = 100$ and an equilibrium angle of $\theta_0{=}\pi$ was also used until the DNA is stretched to 90% of its total contour length, after which it was switched off. Figure 3.1 depicts a portion of the bead-spring DNA chain used in our simulation.

Figure 3.1: Snapshot of a section of extended bead-spring DNA polymer.

To mimic the single molecule experiments discussed in Chapter 1, the stretched DNA was aligned in the x-direction along a two-dimensional surface. Specifically, the target DNA chain was located just above the $x-y$ plane defining the bottom wall of the simulation box with dimensions x–y–z = $100\sigma \times 12\sigma \times 15\sigma$. The DNA ends were immobilized on the surface, while the other beads of the DNA chain were free to fluctuate due to thermal forces. Periodic boundary condition were imposed in the x and y directions, while the upper and lower walls perpendicular to the z direction were repulsive.

The diffusing protein molecules were spherical beads of size $D = 2\sigma$ that interact with the DNA beads via a Lennard-Jones potential of the form given in Eq. 2.3 with $\epsilon \geq k_B T$ and an inter-bead cutoff radius $r_c = 2.5\sigma_{average} = 3.75\sigma$, where $\sigma_{average} = 1.5\sigma$ is the

Figure 3.2: Beads of size 2σ and a bead spring polymer with beads of size σ represent the protein molecules and DNA, respectively.

average of the DNA monomer size σ and the protein molecule size 2σ. Figure 3.2 shows a sample of a short section of DNA chain with several associated protein beads. A repulsive pair interaction between protein beads was imposed using the Lennard-Jones potential with $\epsilon = 4k_BT$ and an inter-bead cutoff radius $r_c = 2.2448\sigma$. These parameters were chosen to insure active excluded volume interactions between protein beads. Prior to each simulation, a collection of N proteins were initially positioned randomly in the simulation box but outside of an exclusion zone of radius 3σ from the DNA chain, and a short simulation run using a soft potential was then used to push off any initially overlapping proteins and create an initial particle configuration file.

3.2 Results

3.2.1 Trajectories of proteins

Protein diffusion along the DNA chain

We first investigated the quasi 1-D Brownian motion of $N = 10$ proteins of size 2σ after they had bound strongly to the stretched DNA. In single molecule experiments, such

one-dimensional diffusion can occur in protein solutions with sufficiently low salt concentration, for which the affinity of proteins for DNA chains is relatively high. We have found in our simulations that a binding energy of $\epsilon = 3k_BT$ is sufficient to achieve protein confinement to the DNA chains for this model system. Figure 3.3 shows the displacements of three of the proteins plotted as function of time. In these conditions, proteins which initially bound to a free segment of the DNA diffused along the chain for a long period of time without dissociation. Notice that, due to the high repulsive interaction between proteins, there are instances of collisions between proteins that restrict their trajectories. Figure 3.4 displays the x, y, and z components of the trajectory of one protein particle diffusing along the DNA chain. Evidently, the protein motion is predominantly along the DNA chain, as indicated by the relatively restricted motion in the transverse (y and z) directions.

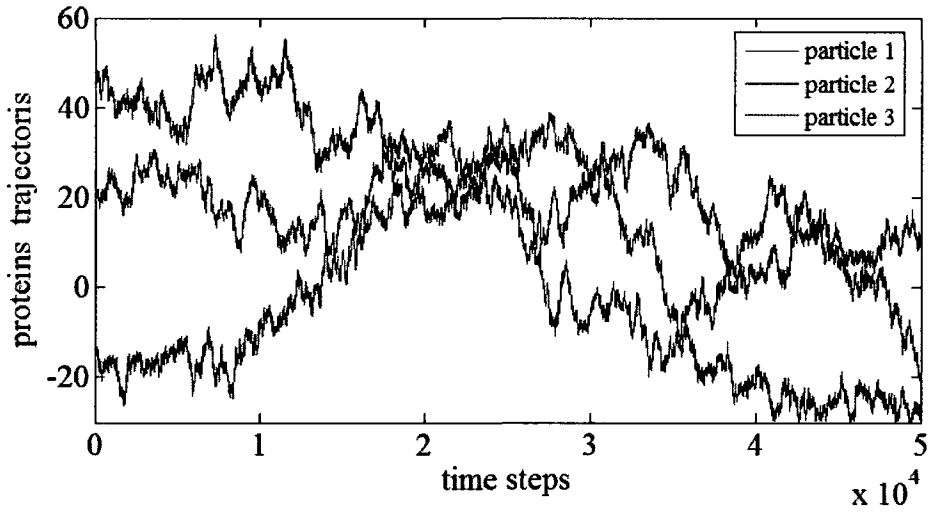


Figure 3.3: Trajectories of three proteins diffusing along a straight DNA by $\epsilon = 3k_BT$.

Figure 3.5 provides a histogram plot of the net displacements of proteins after 5000 time steps, showing that the displacement is distributed symmetrically around zero, as expected for one-dimensional random walk behaviour.

The data shown in Figures 3.3-3.5 for the trajectories of bound proteins along a DNA are qualitatively similar to that reported in several single molecule experiments [25, 47].

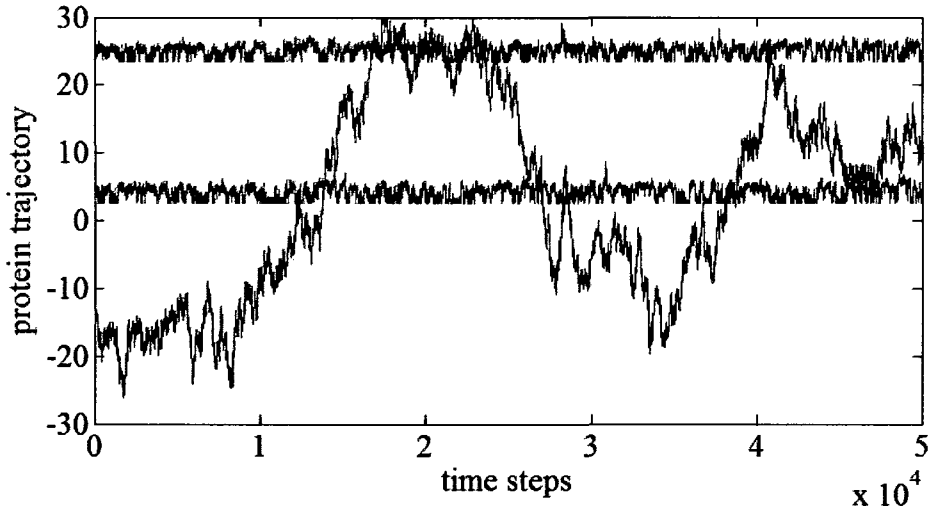


Figure 3.4: Trajectory of a protein molecule along x(red line), y(green line), z(blue line) directions as a function of time steps.

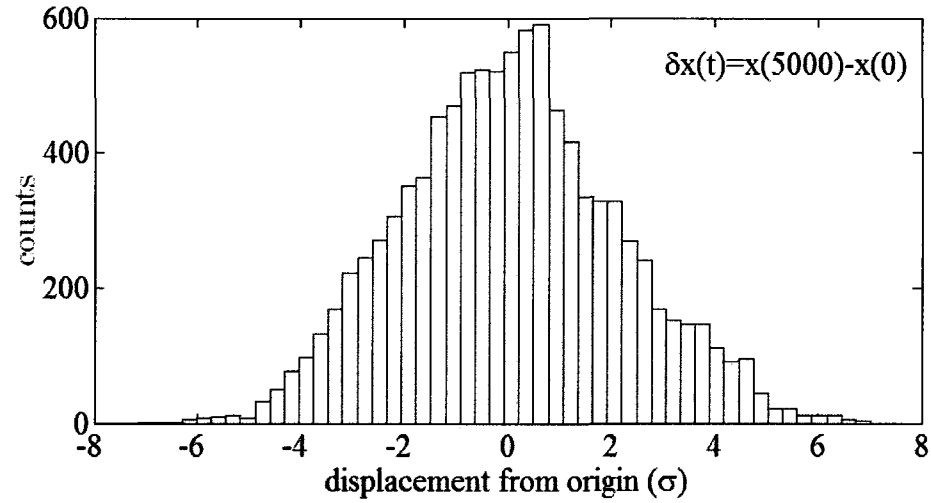


Figure 3.5: Histogram of net displacement after 5000 time steps with the mean value of 0.0520σ.

In order to quantify this behavior and its dependence on binding strength, simulation data were analyzed for protein mean square displacement, the time-course of protein residence on the chain, and the temporal auto-correlation of the protein binding state, as described in the following.

Mean Square Displacement

The 1-D Mean Square Displacement (MSD) of proteins as a function of time was calculated as a time-average over the reference time t_0 from time-series data using

$$\left\langle \Delta r^2(t) \right\rangle = \left\langle [x(t+t_0) - x(t_0)]^2 \right\rangle_{t_0} \tag{3.1}$$

This expression assumes motion exclusively along the x-direction, a limit only valid for sufficiently large ϵ. Later in the thesis this will be properly generalized to 3-D behaviuor. In principle, one expects asymptotic behavior of the form $\langle \Delta r^2(t) \rangle = 2D_{1d}t$ for $t \gg 1$. Figure 3.6 shows $\langle \Delta r^2(t) \rangle$ for three bound protein particles with $\epsilon = 3k_BT$ in a log-log representation. The total number of time steps of the simulation run was 2×10^7, but the mean square displacement was calculated every 50 timesteps (with 1 time step=0.005 τ). Note that the MSD for the particles are virtually indistinguishable. Each MSD exhibits a sub-diffusive regime at early times as a result of the crossover from ballistic to diffusive dynamics with increasing time delay. Figure 3.7 shows a fit of the MSD at long time delays to a power law, $\langle \Delta r^2(t) \rangle \sim t^\alpha$ with $\alpha = 0.97$. Note that the data in the figure indicates that the protein translates by half of its size (σ) within about 4000 time-steps.

Protein residence history

The efficiency of the DNA target search by a protein, which can be measured by the number of new sites visited during a certain search time, is expected to strongly depend on the strength of the DNA-protein binding energy through the combined effects of protein sliding, hopping, and 3-D diffusion. One simple means of characterizing these processes is through an analysis of the time history of bound and unbound states. For such a study, simulations were conducted as a function of binding energy ϵ for $N = 5$ protein particles and 2×10^7 time steps. Every 10 time-steps, the distance d_{pr} of each protein relative to the DNA was determined and the result was tabulated as either a bound (1) or unbound (0) state according to an ad hoc distance cut-off, chosen to be $d_{bd} = 2\sigma$. Thus, if $d_{pr} \leq d_{bd}$, the protein was recorded as bound (on=1) to the DNA at that timestep, otherwise an unbound

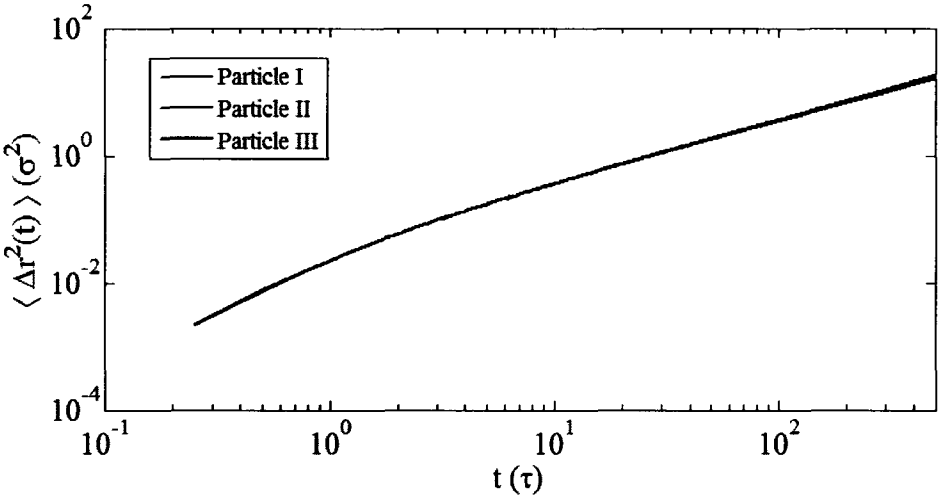


Figure 3.6: Log-Log plot of mean square displacement of three proteins along the DNA axis as a function of time.

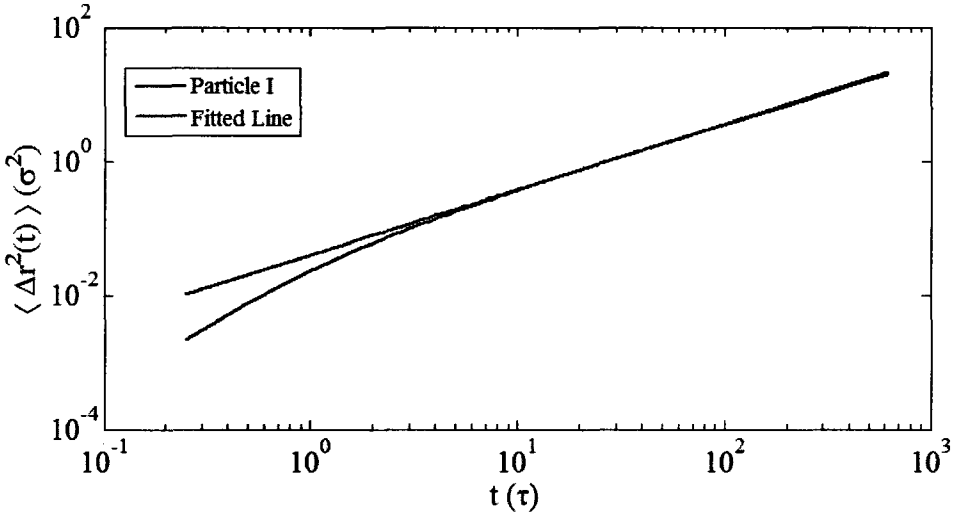


Figure 3.7: Log-Log plot of mean square displacement of a protein along DNA axis as a function of time, along with a power law fit, $\langle \Delta r^2(t) \rangle \sim t^{\alpha}$ with $\alpha \simeq 0.97$, at long times.

state was recorded (off=0). The precise choice of cut-off distance d_{bd} did not qualitatively change the results. Intuitively, one might expect that weakly interacting proteins (with small ϵ) are rarely found in the bound state, while strongly interacting proteins (with large ϵ) remain primarily in the bound state once they have first encountered the DNA chain.

Figure 3.8 shows a time-series plot of bound (occupation state=1) and unbound (occupation state=0) proteins for the case of $\epsilon = 3k_BT$, a relatively strong binding energy. The figure indicates that proteins with $\epsilon = 3k_BT$ spend the majority of time bound to the DNA (executing sliding motion) and make short excursions off the DNA chain (hops), in agreement with expectation.

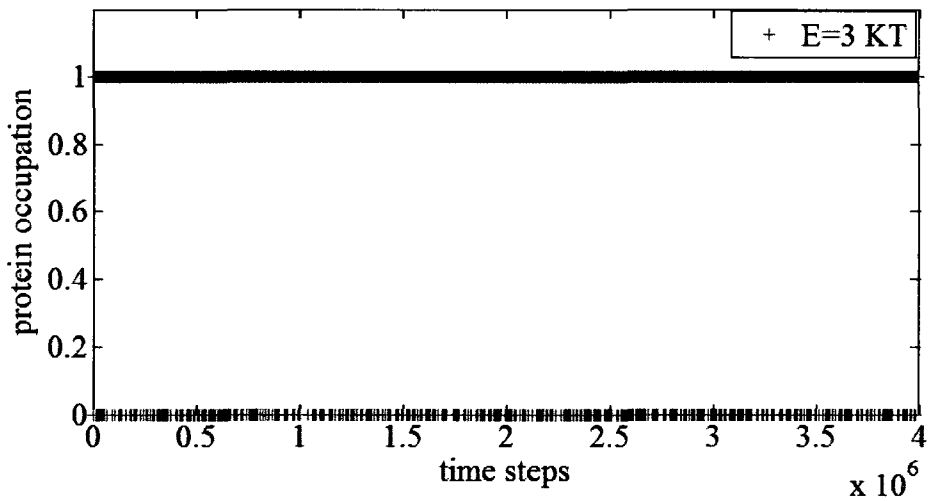


Figure 3.8: Protein occupation on the DNA chain at different time-steps for $\epsilon = 3k_BT$.

On the other hand, Figure 3.9 shows an analogous time-series plot of bound and unbound proteins for the case of $\epsilon = 1k_BT$, a relatively weak binding energy. The figure indicates such proteins spend the majority of time unbound to the DNA (executing 3-D diffusion) and make only temporary visits to the DNA chain, also as expected.

At intermediate binding energies, proteins exhibited a balance of 1-D sliding, hopping, and 3-D diffusion. Figure 3.10 shows a time-series plot of bound and unbound proteins for the case of $\epsilon = 2k_BT$. Note that in this case there are fewer of the small scale hops observed in stronger binding conditions (so called micro-hops) and more larger-scale macro-hops.

The above results are consistent with the work presented in Ref. [39], which advocated a scenario with a shifting balance between DNA sliding, hopping and 3-D dissociation that depends on the salt concentration conditions (a known modulator of protein binding). As will be discussed below, the intermediate binding energy conditions are most conductive to

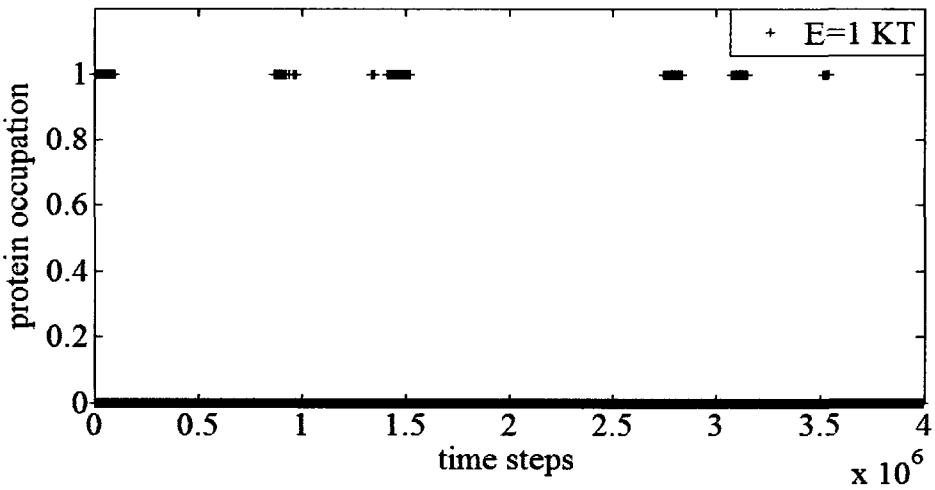


Figure 3.9: Protein occupation on the DNA chain at different time-steps for $\epsilon = 1k_BT$.

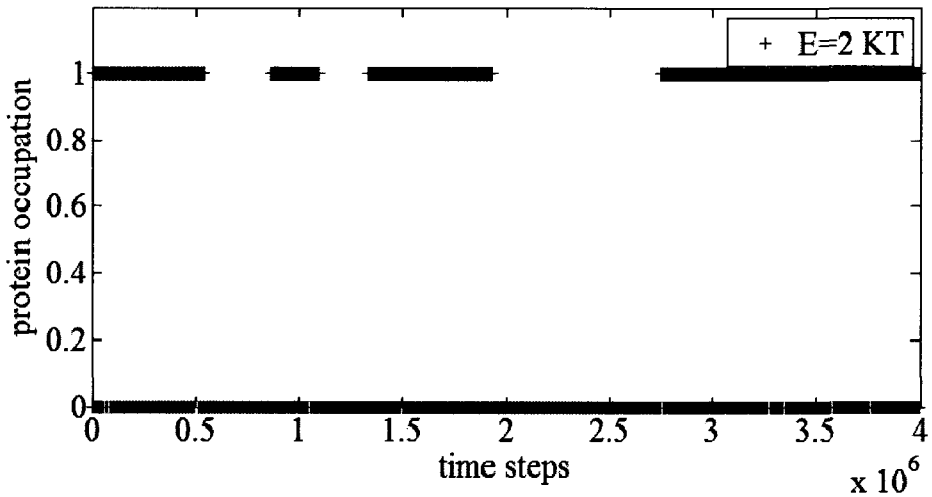


Figure 3.10: Protein occupation on the DNA chain at different time-steps for $\epsilon = 2k_BT$.

efficient DNA target discovery by proteins, as these conditions lead to a balance between 1-D and 3-D diffusive motions [28, 40, 41].

Correlation of protein binding

Using the above time-series data for bound and unbound protein states, the temporal autocorrelation between states was studied. Here we focus on the case of the autocorrelation function $g(t)$ for bound states with different binding energies, defined by $g(t)=<$

$on(t + t_0) \cdot on(t_0) >_{t_0}$, with a time-average over reference times t_0. This represents the probability of finding a protein on the chain at time $t + t_0$ if it was on the chain at an earlier time t_0. The times were chosen in increments of 10 time steps and thus $g(t)$ was averaged over 2×10^6 events. The un-normalized correlation between bound protein events is plotted versus time in Figure 3.11 for different binding energies, $\epsilon = 1, 2, 3k_BT$.

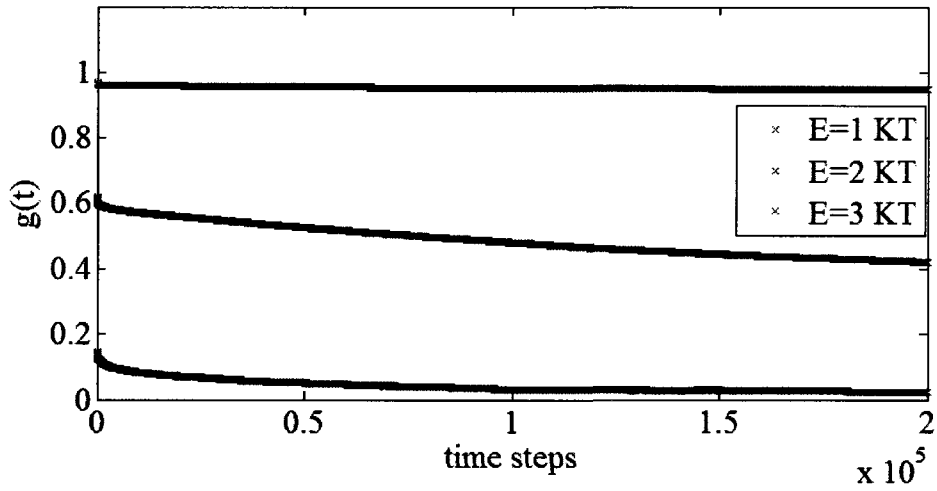


Figure 3.11: Temporal autocorrelation function of bound protein for $\epsilon = 1, 2, 3k_BT$.

Normalized g(t) were then obtained by dividing the autocorrelation function by the total fraction of on states (given by total number of time steps that the protein is on the chain over the total number of protein states (i.e. time steps)). Table 3.1 provides normalization data for the cases studied.

Binding Energy	$\epsilon = 1k_BT$	$\epsilon = 2k_BT$	$\epsilon = 3k_BT$
Total number of on states	300076	1250107	1948280
Total number of states	2×10^6	2×10^6	2×10^6
Normalization factor	0.1500	0.6250	0.9741

Table 3.1: $g(t)$ normalization data for three different binding energies.

Figure 3.12 shows that while bound states for high binding energies exhibit long-time

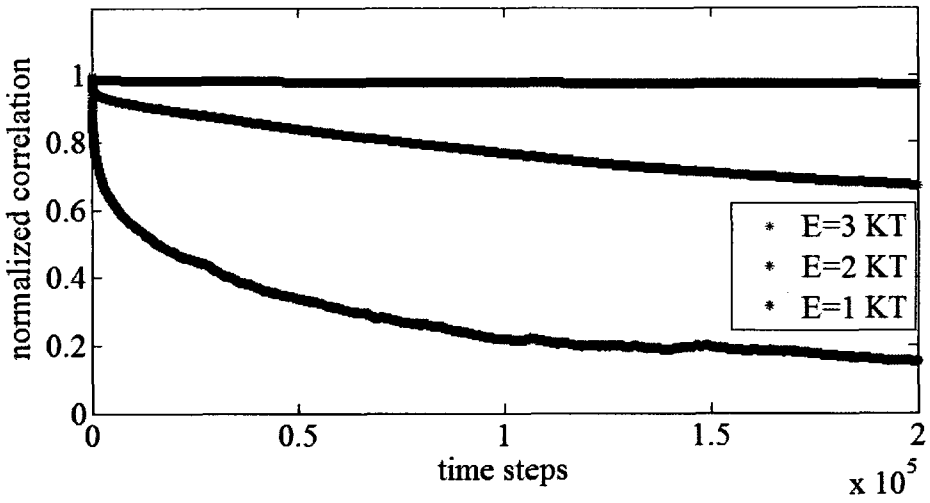


Figure 3.12: Normalized temporal autocorrelation function for $\epsilon = 1, 2, 3k_BT$.

correlation (indicative of 1-D sliding motion), as the binding energy decreases the bound states become more rapidly decorrelated in time, as expected for proteins that execute 3-D diffusive excursions off the DNA chain. For instance, normalized g(t) decays to 50% of its value after 10,000 time stepsfor $\epsilon = 1k_BT$,where for large ϵ it is much slower.

3.2.2 Mean First Passage Time (MFPT)

The mean first passage time (MFPT) is the average time required for a protein to find its specific target on a DNA contour. This is the most important quantity for characterizing target search processes for DNA-protein interactions. A series of simulation studies were performed to elucidate the dependence of bulk protein concentration and protein-DNA binding affinity on the MFPT. First, the MFPT was investigated in the case of pure 1-D motion of a strongly bound protein on an extended DNA chain. For the studies of extended DNA systems, the end bead on one end of the DNA chain was designated as the specific target site and a collection of N proteins ($N = 5, 10, 20$) are randomly placed on an extended DNA chain. The simulations involved many repetitions of a basic first-passage time (FPT) simulation, as follows. For a given number of bound proteins, a simulation was run for a total time sufficient for target discovery by at least one of the proteins. This basic

FPT simulation was repeated at least 10,000 times, each starting with a new set of initial protein configurations on the chain. In post-simulation analysis, the first instance of a protein finding the target DNA site was identified for each run and the elapsed time from the start of the simulation to this target discovery event was recorded as the FPT. The MFPT was obtained from the mean value of the FPT from the entire sequence of runs.

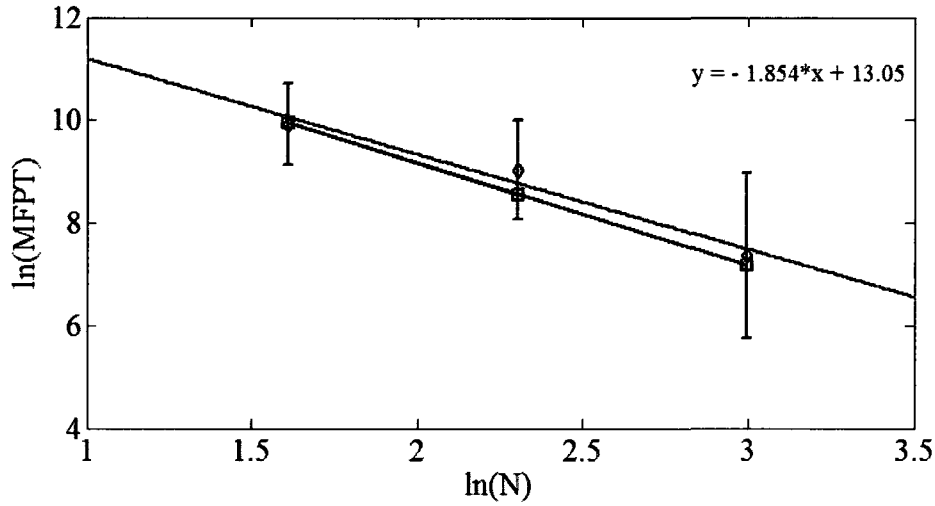


Figure 3.13: Log-log plot of the MFPT vs. protein number N on the chain for 10,000 events at $\epsilon = 3k_BT$ (blue diamonds). Error bars shown are the standard deviations. The red line is a fit to the data, giving MFPT$\sim N^{-1.85}$. For comparison, the inverse-squared dependency predicted by theory, MFPT$\sim N^{-2}$, is shown in green squares connected by black line.

Figure 3.13 shows the dependence of the MFPT on the number of bound proteins for this high binding energy study. In this case, the MFPT required for a protein to find its target by pure sliding was found to decrease as a power-law of the number of bound proteins MFPT$\sim N^{-1.85}$.this is consistent with other experimental and computational studies [35,19]. For instance, the simulation is in close agreement with the theoretical result for pure 1-D diffusion of self-avoiding proteins predicted by Ref. [37] that MFPT$\sim C^{-2}$, where $C \propto N$ is the one-dimensional protein concentration on a structureless, linear DNA chain. The discrepancy between the simulation and theoretical exponents may be due to the

different consequences of self-avoidance in these two studies. While the self-avoidance in the 1-D theoretical model strictly rules out the displacement of one protein passed another, the repulsive interactions between proteins do not prohibit such relative motion in the 3-D embedding environment of simulation, since two proteins may pass each other on opposite sides of the chain. This would act to soften the dependence of the MFPT on N.

In all experimental studies, the stretched DNA chain with a target site is exposed to proteins in a bulk 3-D solution. Thus, the target search is usually a sequential process in which proteins must first diffuse in 3-D to find the DNA substrate before executing a 1-D sliding diffusive search for the target site. Therefore to conduct a more realistic simulation, we next randomly positioned the proteins inside the simulation box and simulated the combined 3-D diffusion and 1-D sliding processes involved in the target search. The protocol for conducting MFPT simulations in 3-D was similar to the quasi-1-D case described above. A series of many short simulations were conducted to obtain independent estimates of the FPT for different initial conditions (in this case, the initial position of the proteins in the simulation cell). In this case, the FPT includes the initial time to find the DNA chain by 3-D diffusion. These FPTs were then averaged to obtain the MFPT. Histograms of the FPT are plotted for 6000 events in Figure 3.14 for two analogous systems with $N = 20$ and $\epsilon = 3k_BT$. Here the difference between the two was just in the initial position of proteins. For the 1-D data, the proteins were initially placed randomly on the DNA chain; while for the 3-D + 1-D data, the initial positions were in the bulk. The histogram data clearly shows that for search processes involving a combination of 1-D and 3-D diffusion of proteins, the FPTs are more widely distributed and posses a larger MFPT than in the case of pure 1-D diffusion.

MFPT and Binding Energy

In this section, the impact of adsorption energy on mean first passage time is presented. For these studies, the protein number (N) and the protein-DNA interaction energy (ϵ) were independently varied in a set of simulations. The proteins were randomly positioned in the simulation box (bulk 3-D scenario) with the protein-DNA interaction energy chosen

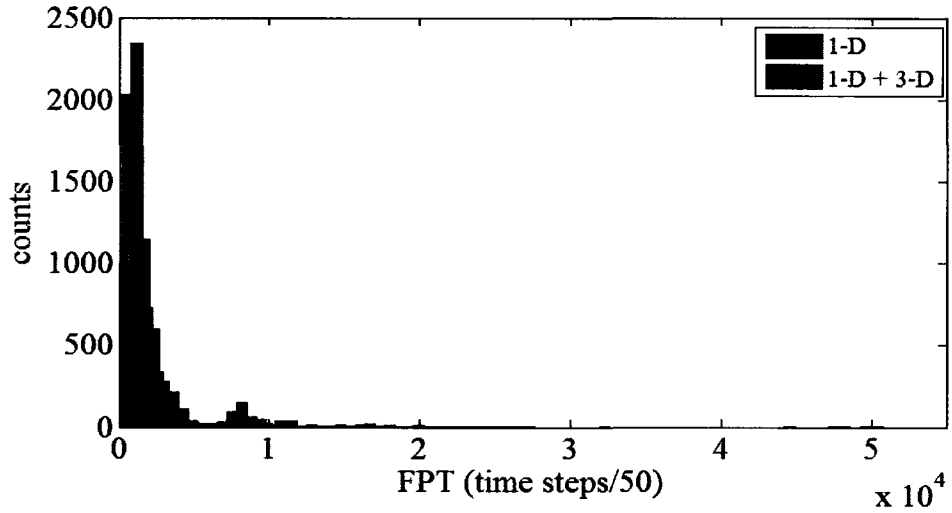


Figure 3.14: Histograms of first passage time (FPT) for 20 proteins diffusing on DNA chains with two different initial conditions: (i) proteins initially placed on the chain (red) and (ii) proteins are initially positioned in the bulk (blue).

from a value in a range from $\epsilon = 1k_BT$ to $5k_BT$ in increments of $0.5k_BT$. As before, FPT simulations were repeated 10 000 times, with different initial protein configuration for each simulation, and the results were used to obtain the MFPT as a function of N and ϵ.

Figure 3.15b shows the MFPT for two protein concentrations ($N = 10$ and $N = 20$) as a function of ϵ . The error bar shown is the standard error of the mean. Note the non-monotonic nature of the MFPT for a constant number of proteins: as the binding energy increased, the MFPT first decreased with ϵ, reaching a minimum value, and finally increased with ϵ. For $N = 10$, the minimum value of the MFPT occured somewhere between $\epsilon = 2k_BT$ and $3k_BT$, after which there is a pronounced increase in MFPT with increasing ϵ. On the other hand, for $N = 20$ the apparent minimum near $\epsilon = 3k_BT$ was quite a bit more shallow and the rise in MFPT at higher ϵ was negligible (the error bars makes it hard to pinpoint the precise location of the minimum binding energy). This is presumably due to crowding effects at linear high protein density dominating the sliding behavior [30, 11].

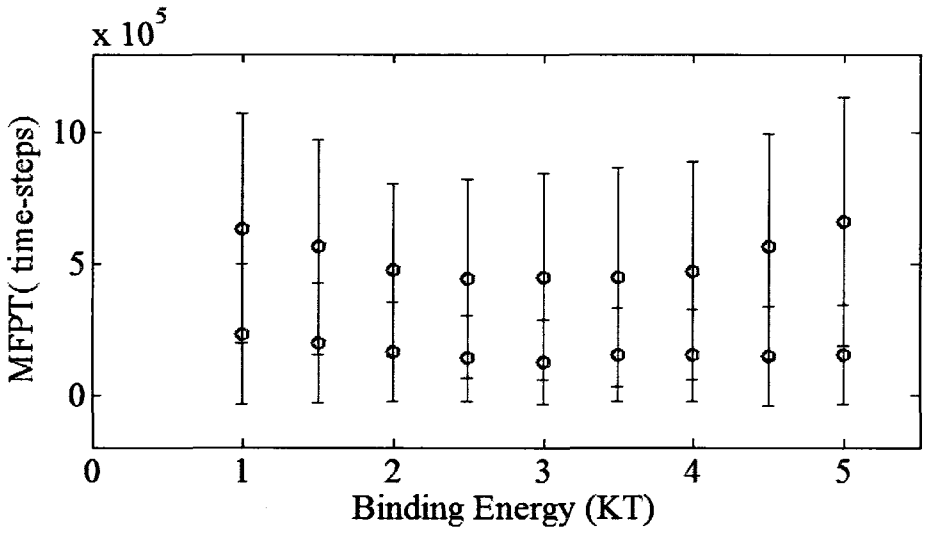
Interestingly, Slutsky and Mirny [28] also noted that the energy landscape of diffusion

along DNA plays an important role in controling the balance between sliding and three dimensional diffusion. They predicted that the protein-DNA binding energy landscape has an optimal value of approximately $2k_BT$ that maximizes the DNA target search rate. Another point of view [41] relates sliding length to target search time. In this picture, as sliding length increases, the target search time decreases until it reaches its minimum value at an optimum sliding length (or binding energy), beyond which it grows again. These feature were captured quantitatively in a stochastic theory given in Ref. [40], which determined the relative search time τ/τ_s, where τ_s is the ordinary diffusion time given by Smoluchowski theory [8], as a function of the dimensionless parameter, $y = \exp(E_{ads}/k_BT)$,

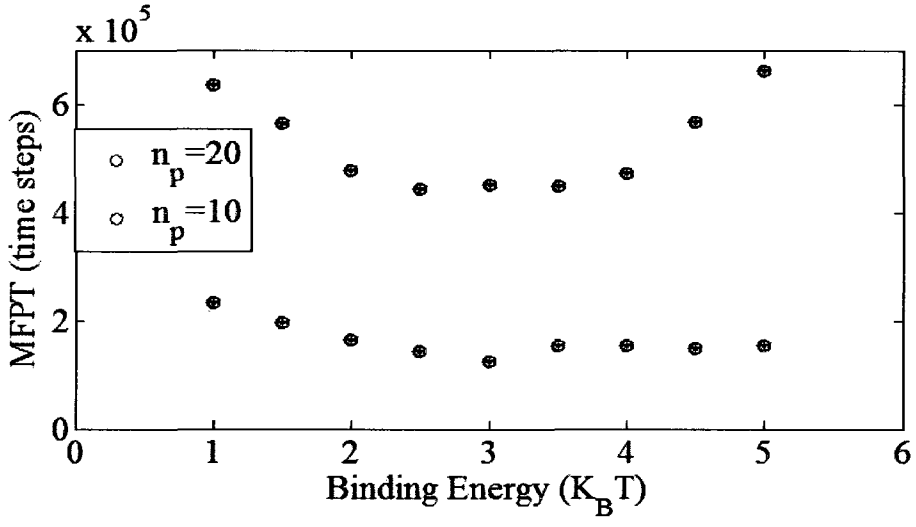
$$\frac{\tau}{\tau_s} = \frac{a}{r}\left(\frac{1}{\sqrt{n_{ads}yd}} + \frac{n_p\sqrt{y}}{n_{ads}^{\frac{3}{2}}\sqrt{d}} + \frac{2}{\sqrt{n_p yd}}\right) \tag{3.2}$$

where E_{ads} is the non-specific protein-DNA binding energy, a is the target site size, n_p is the number of proteins in the solution, n_{ads} is the number of adsorbed proteins on DNA, and d is the ratio of 1-D and 3-D diffusion coefficients $d = D_1/D_3$. According to Eq.3.2, the relative search time has a complex dependency on non-specific binding energy, which is plotted in Figure 3.16 for a set of specified parameter values. However, we note the qualitative agreement of the plots shown in Figures 3.15 and 3.16. As depicted in the figure, the search time is initially a decreasing function of the adsorption energy, reaching the optimal value (at about $4k_BT$) before increasing with adsorption energy beyond the optimal value.

We note here that the standard deviation of the MFPT, shown for instance in Figure 3.13, is always rather large, although the mean value itself does not change significantly and the standard deviation of the mean is always a small value. There is some theoretical justification for the inherent noise observed in our simulation data. Refs. [48] and [49] argue that the MFPT is an inherently noisy quantity due to its dependence on the random motion of a finite collection proteins in space, with some possible reduction of noise for cases of enhanced one dimensional diffusion (sliding). This is consistent with what we



(a)



(b)

Figure 3.15: MFPT versus binding energy for N = 20 (red circles) and N = 10 (blue circles). In a, the error bar shown is the standard deviation of the data; in b, the error bar is the standard error of the mean.

have seen in our results: large standard deviations of the MFPT with very modest reduction at high binding energies which favor sliding processes.

Histogram of FPTs

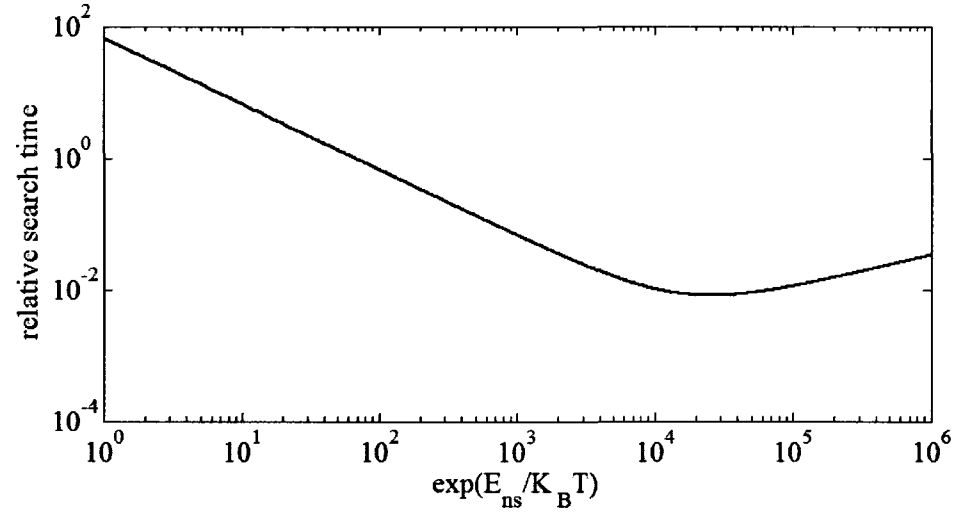


Figure 3.16: Relative search time as a function of the dimensionless adsorption strength for $a = 1$ nm, $r = 30$ nm, $n_{ads} = 1000$, $n_p = 1$, and $d = 0.001$.

The inherent noisy behavior of the target search process is most evident in the distribution of FPTs. Figure 3.17 shows a sample histogram of FPTs for the case of diffusing 20 proteins for three different binding energies ($\epsilon = 1, 2, 3k_BT$). Note the noisy data and, in particular, the gaps that occur for low binding energies.

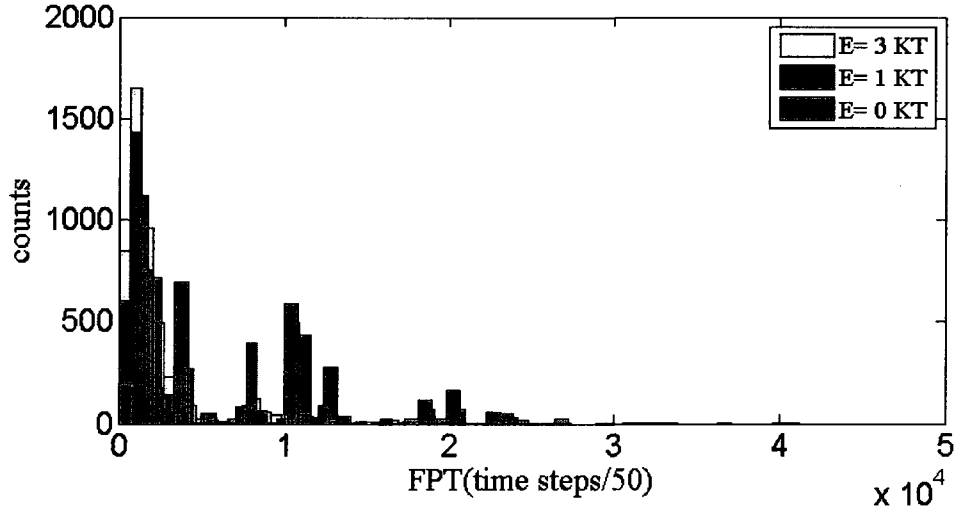


Figure 3.17: Histograms of mean first passage time of 20 diffusing proteins with different binding energies.

This noisy behaviour is especially prominent at low volume fraction of proteins in the bulk, a limit often relevant to experiment. The effective protein volume fraction for the configuration of Figure 3.17 is $\phi \simeq 3.7\%$. Additional simulations on more concentrated systems were conducted to explicitly demonstrate the effect of concentration on the noisy behavior of the FPTs. Figure 3.18 and 3.19 show a histogram constructed for the case of 100 diffusing proteins (corresponding to an effective volume fraction of $\phi = 16\%$) with a weak affinity for the DNA chain ($\epsilon = 1k_BT$) and at a high binding energy ($\epsilon = 3k_BT$), respectively. Clearly in this case the noise was substantially reduced. However, such large protein volume fractions are not experimentally relevant and may also result in saturation of the DNA with proteins at high ϵ, a different physical scenario than what we had chosen as the focus of this study.

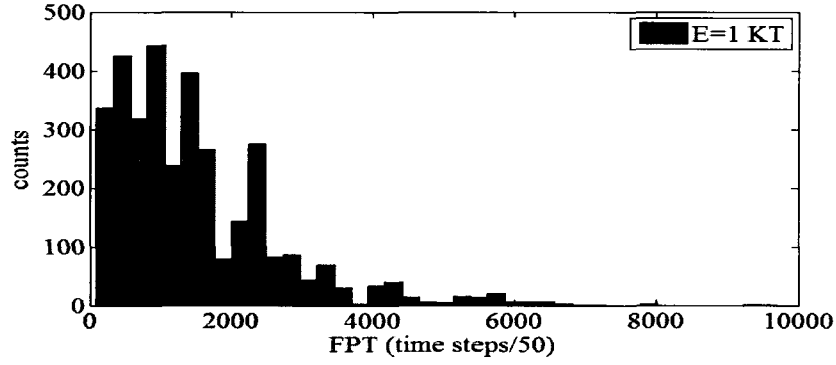


Figure 3.18: Histogram of mean first passage time for $n_p = 100$ and $\epsilon = 1k_BT$.

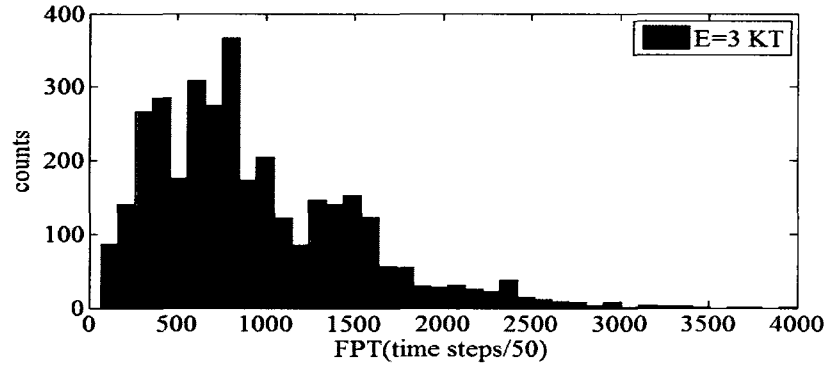


Figure 3.19: Histogram of mean first passage time for $n_p = 100$ and $\epsilon = 3k_BT$.

3.3 Conclusion

In this chapter we used brownian dynamics simulation to characterize the process of proteins searching to recognize a specific target on a DNA chain. The protein molecules were simply modeled as single beads of size 2σ and the DNA was represented as an extended bead-spring polymer. The interactions between proteins and DNA monomers were defined through the attractive Lennard-Jones potential with the energy depth ϵ. The efficiency of DNA target search by proteins highly depended on the strength of DNA-protein binding. For high binding energy, $\epsilon = 3k_BT$, the proteins spent most of their time sliding on DNA without dissociation. Through this quasi 1-D diffusive process, the proteins were able to scan the sites on DNA to find the specific cognate site. As the binding energy decreased, the hopping mechanism (micro- and macro-hops), whereby the proteins executed limited 3-D diffusive excursions off the DNA chain, was seen to play a more significant role in the search process. However, at sufficiently low binding energy ($\epsilon = k_BT$) the proteins were found to be most frequently detached from the DNA and undertaking a three-dimensional diffusive search of the bulk. The mean first passage time (MFPT), the mean time required for first discovery of the DNA target site, was found to depend on both protein concentration and the binding energy between proteins and DNA. Interestingly, the dependence of the MFPT on binding energy was not monotonic. With increasing binding energy, one dimensional diffusion became increasingly dominant over three dimensional diffusion, with proteins increasingly confined to the immediate vicinity of chain for extended periods of time. The time required for a given protein to find its specific target on the DNA initially decreased when the binding energy increased. However, once this binding energy was sufficiently large, the proteins spent more of their time oversampling small regions of DNA, thereby slowing the sliding process. Thus, the existence of an optimal binding energy was verified for minimizing the target search time, in accord with previous theoretical predictions[28, 40, 41]. In this study, the long-range inter-segmental transfer of proteins was absent due to the extended nature of the DNA chains. Therefore, the search process was limited to three mechanisms: 1-D sliding, 3-D bulk diffusion, and local hopping. Each

of these were shown to be operative in the protein dynamics and the target search process. The optimal binding energy was found to be near $\epsilon = 3k_BT$, for which the proteins rarely detached from the chain. Thus, the most efficient target search process was dominated by 1-D sliding mode diffusion. This result differs from the studies of Ref.[39], for which optimal search efficincy was predicted to occur when sliding was only 20% of the total search. This difference could easily arise from the difference in the models for the DNA (helical vs linear bead-spring) and its interactions with the proteins (LJ vs electrostatic).

Protein concentration was also found to be an important parameter in controlling the target search process. For high protein concentration, there were always proteins in the solution close to the target that could efficiently reach it through ordinary three dimensional diffusion. In contrast, for low protein concentration, the diffusing proteins first explored the bulk (for a long time) to find the DNA chain and subsequently executed a sliding search along a finite length of the DNA to find the target. Such behavior was observed experimentally (e.g. see Figure 4 of Ref. [26]) and the protein concentration effect on the specific target rate was also studied in Ref. [40]. In this theoretical model, at low protein concentration the protein molecule is predicted to spent most of its time on binding and unbinding events, limiting the sliding length along the DNA between pairs of binding and unbinding events, and thereby increasing the total search time. In our simulations at low protein concentrations, long sliding runs occurred at high binding energy after the protein found the DNA chain, facilitating the target search process. Our simulation studies of coase-grained models were seen to reproduce many of the salient features of previous experimental and theoretical studies, and provided comple- mentary information not yet available from other approaches. In particular, we have shown that a simulation study can provide insight into the detailed nature of protein motion on and in the vicinity of a DNA chain, and the potential correlated motion of proteins. Such information is not easily obtained from theoretical kinetics models and single molecule ex- periments. In the next chapter of this thesis the constraint of extended DNA chain confor-Results 49 mation is relaxed, allowing studies of the effects of DNA conformation on DNA-facilitated

protein transport and the target search process In this chapter we used Brownian dynamics simulation to charactrize the process of proteins searching to recognize a specific target on a DNA chain. The protein molecules were simply modeled as a single beads of size 2σ and the DNA was represented as an extended bead-spring polymer. The interactions between proteins and DNA monomers were defined through the attractive Lennard-Jones potential with the energy depth ϵ. The efficiency of DNA target search by proteins highly depended on the strength of DNA-protein binding. For high binding energy, $\epsilon = 3k_BT$, the proteins spent most of their time sliding on DNA without dissociation. Through this quasi 1-D diffusive process, the proteins were able to scan the sites on DNA to find the specific cognate site. As the binding energy decreased, the hopping mechanism (micro- and macro-hops), whereby the proteins executed limited 3-D diffusive excursions off the DNA chain, was seen to play a more significant role in the search process. However, at sufficiently low binding energy ($\epsilon = k_BT$) the proteins were found to be most frequently detached from the DNA and undertaking a three-dimensional diffusive search of the bulk.

The mean first passage time (MFPT), the mean time required for first discovery of the DNA target site, was found to depend on both protein concentration and the binding energy between proteins and DNA. Interestingly, the dependence of the MFPT on binding energy was not monotonic. With increasing binding energy, one dimensional diffusion became increasingly dominant over three-dimensional diffusion, with proteins increasingly confined to the immediate vicinity of chain for extended periods of time. The time required for a given protein to find its specific target on the DNA initially decreased when the binding energy increased. However, once this binding energy was sufficiently large, the proteins spent more of their time oversampling small regions of DNA, thereby slowing the sliding process. Thus, the existence of an optimal binding energy was verified for minimizing the target search time, in accord with previous theoretical predictions[28, 40, 41]. In this study, the long-range inter-segmental transfer of proteins was absent due to the extended nature of the DNA chains. Therefore, the search process was limited to three mechanisms: 1-D sliding, 3-D bulk diffusion, and local hopping. Each of these were shown to be operative in

the protein dynamics and the target search process. The optimal binding energy was found to be near $\epsilon = 3k_BT$, for which the proteins rarely detached from the chain. Thus, the most efficient target search process was dominated by 1-D sliding mode diffusion. This result differs from the studies of Ref.[39], for which optimal search efficincy was predicted to occur when sliding was only 20% of the total search. This difference could easily arise from the difference in the models for the DNA (helical vs linear bead-spring) and its interactions with the proteins (LJ vs electrostatic).

Protein concentration was also found to be an important parameter in controlling the target search process. For high protein concentration, there were always proteins in the solution close to the target that could efficiently find it through ordinary three-dimensional diffusion. In contrast, for low protein concentration, the diffusing proteins first explored the bulk (for a long time) to find the DNA chain and subsequently executed a sliding search along a finite length of the DNA to find the target. Such behaviour was observed experimentally (e.g. see Figure 4 of Ref. [26]) and the protein concentration effect on the specific target rate was also studied in Ref. [40]. In this theoritical model, at low protein concentration the protein molecule is predicted to spent most of its time on binding and unbinidng events, limiting the sliding length along the DNA between pairs of binding and unbinding events, and thereby increasing the total search time. Whereas, in our simulations at low protein concentrations, long sliding runs occurred at high binding energy after the protein found the DNA chain, facilitating the target search process.

Our simulation studies of coase-grained models were seen to reproduce many of the salient features of previous experimental and theoretical studies, and provided complementary information not yet available from other approaches. In particular, we have shown that a simulation study can provide insight into the detailed nature of protein motion on and in the vacinity of a DNA chain, and the potential correlated motion of proteins. Such information is not easily obtained from theoretical kinetics models and single molecule experiments. In the next chapter of this thesis the constraint of extended DNA chain conformation is relaxed, allowing studies of the effects of DNA conformation on DNA-facilitated

protein transport and the target search process.

Chapter 4

Proteins and Coiled DNA

DNA molecules found *in vivo* are almost always sufficiently long that they adopt coil-like conformations. In the coiled state, distant base pairs in the sequence can be found temporarily in close spatial proximity, providing opportunities for rapid transport of non-specifically bound protein to distant sections of the chain through inter-segmental transfer events. Such events are distinct from the 1-D sliding, local hopping, and 3-D diffusive motion considered in the last chapter for extended DNA chains, for which inter-segmental transfer is not relevant. For coiled DNA, inter-segmental transfer has been proposed as an important mechanism for more efficient target searching by DNA-associated proteins [11, 35]. However, experimental single molecule studies of the effects of DNA conformation on inter-segmental transfer of proteins are especially challenging, due to the difficulty in visualization of small-scale protein motion. An attractive alternative is the use of computer simulation to characterize these phenomena. This chapter presents studies of the diffusion of proteins along and within a coiled DNA chain using Brownian dynamics simulation. After characterization of the association of diffusing proteins with the DNA coil and their subsequent motion along the DNA contour, this study focused on determination of the MFPT for a protein initially in solution to find a specific target on the DNA, and the role of inter-segmental transfer in this process.

4.1 Model

As in the previous chapter, the DNA molecule was modeled as a linear bead spring polymer comprised of N beads of size σ joined together by a modified FENE bond potential of the form given in Eq.2.4 with spring constant $K = 100\epsilon/\sigma^2$. The strong spring constant restricted the inter-bead bond length to $R < R_0 = 1.22\sigma$. The repulsive part of the bond potential had cutoff radius of $r_c = 1.1224\sigma$ to prevent particle overlap. In order to allow for a sufficiently coiled conformation, a relatively large polymerization index, $N = 500$, was chosen. (i.e. a total of 7500 base pairs). The DNA chain was inserted into a simulation box of size $L_x \times L_y \times L_z = 50 \times 30 \times 32\sigma^3$ with periodic boundary condition imposed on all dimensions of the simulation box. The coiled state of the DNA was obtained by relaxing the initial chain configuration during a prolonged simulation at finite temperature. Once the radius of gyration of the coil reached a steady state, the end monomers were fixed in position (with end-to-end separation of 5σ). This coil configuration served as the initial state for the first set of DNA-protein simulations. Analysis of the coil conformation shows the coil persistent length is 6 beads equals to 90 base pairs.

In order to prevent protein-mediated bridging of distal monomers of the DNA chain (by shared bonding of DNA monomers to a common protein), a new coarse-grained model protein was utilized. This model protein is composed of compact aggregates of 6 bonded beads, each of size σ. Only one of these beads (the head) possessed an attractive LJ potential for DNA monomers (with an interaction parameter ϵ in the range $k_BT \leq \epsilon \leq 4k_BT$ and a cutoff radius between DNA monomers and the protein head of $r_c = 3\sigma$). The remaining 5 beads served to prevent the head bead from interacting simultaneously with more than one DNA monomer. They were connected to the head bead and to each other by a strong harmonic potential with spring constant $\kappa = 500\epsilon/\sigma^2$ and equilibrium bond length $r_0 = 0.92\sigma$.

The non-head beads also were given a short-range repulsive interaction with the DNA monomers of $\epsilon = k_BT$ and cutoff radius $r_c = 1.1224\sigma$ as well as repulsive interactions

between proteins. Figure 4.1 shows a snapshot of these proteins interacting with a coiled DNA molecule.

Figure 4.1: Snapshot of a 500-monomer DNA coil (red and blue beads) in the presence of globular proteins (orange molecules with grey head beads). The central target bead and end monomers of the DNA chain are blue.

4.2 Results

4.2.1 Trajectories of proteins

Protein Mean Square Displacement

As in the previous chapter, we first investigated diffusion of tightly bound proteins along the DNA chain. To do so, we placed a single protein randomly on one of the DNA sites with a binding energy of $\epsilon = 4k_BT$ and simulated its motion along the DNA coil for 3×10^7 time steps, recording particle coordinates every 20 time steps . The 3-D mean square displacement of the protein was calculated as a time-average over the reference

time t_0 from this time-series data using

$$\left\langle \Delta r^2(t) \right\rangle = \left\langle [\vec{r}(t+t_0) - \vec{r}(t_0)]^2 \right\rangle_{t_0} \tag{4.1}$$

where $\vec{r} \equiv (x, y, z)$ is the 3-D position vector of the center of the protein head monomer. As with the case of strong binding of proteins to extended chains studied in the last chapter, the protein motion is confined to the vicinity of the DNA chain and therefore the protein executes long-lived, quasi 1-D (curvilinear) Brownian motion *along* the DNA chain. However, since the DNA chain trajectory is that of a coil embedded in three dimensions, the protein MSD also has a 3-D character, albeit with a sub-diffusive exponent that is determined by the coil geometry. Figure 4.2 shows a log-log plot of the protein MSD vs. time for this strongly bound case of $\epsilon = 4k_BT$. Here $\langle \Delta r^2(t) \rangle \sim t^\alpha$ at long times, with $\alpha \simeq 0.85$. If the coil was a true Gaussian coil, we would expect $\alpha \simeq 0.5$. However, our finite length DNA coil is significantly more swollen than a Gaussian coil. Moreover, subsequent analysis of time series data showed that the proteins make macro-hops between distant sections of the DNA chain when these are in close proximity. Such hopping also modifies the nature of the protein transport, as will be discussed below.

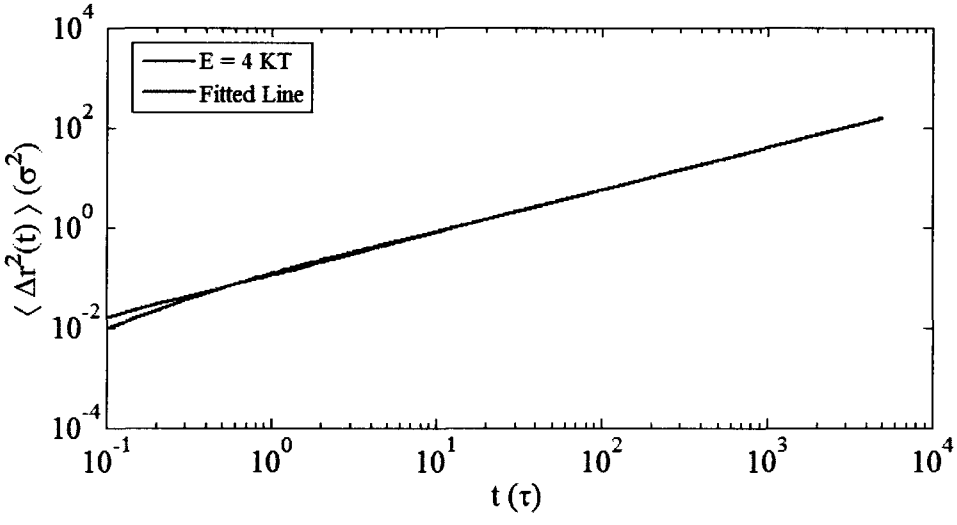


Figure 4.2: Mean square displacement of a protein molecule along DNA coil as a function of time for $\epsilon = 4KT$, along with a power law fit, $\langle \Delta r^2(t) \rangle \sim t^\alpha$ with $\alpha \simeq 0.85$

At sufficiently low binding energies, pure 3-D diffusive motion of the protein should be observed. Figure 4.3 shows a log-log plot of the protein MSD vs. time for the case of $\epsilon = k_B T$, a weak binding energy that allows proteins easy escape from the DNA chain. As expected, simple diffusive behaviour of proteins is observed for time delays beyond approximately $t = 20\tau$. Note, however, that the effective 3-D diffusion constant is for the protein model used in the chapter, composed of 6 bound LJ beads of size σ, and this is different from that of the single bead model in the previous chapter. As will be demonstrated later, the binding affinities of the two beads are also different.

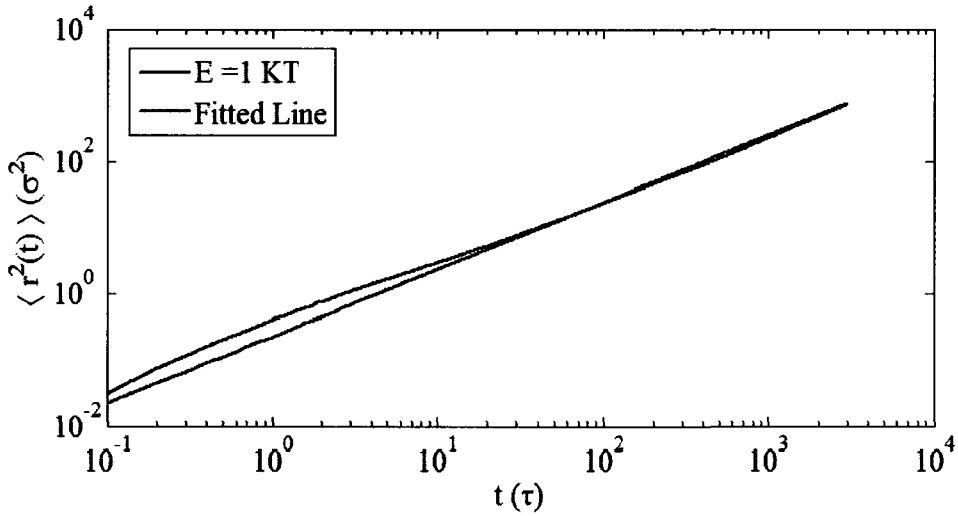


Figure 4.3: Mean square displacement of a protein molecule along DNA coil as a function of time for $\epsilon = 1KT$, along with a power law fit, $\langle \Delta r^2(t) \rangle \sim t^{\alpha}$ with $\alpha \simeq 1.0$ at long times.

Protein residence history

In general, we expect a transition from curvilinear diffusion of proteins along the DNA chain to simple 3-D diffusion of proteins in the box as the binding energy is decreased from high to low levels. Next we investigated the dependence of the relative fraction of sliding, hopping and 3-D diffusion search mechanisms on protein-DNA binding energy. To do so, we used the time analysis methods described in the previous chapter to characterize the history of bound and unbound states for a collection of 10 protein molecules as a function of

binding energy in the range $\epsilon = 1{-}3k_BT$. These proteins were positioned randomly within the simulation box but outside an exclusion zone of distance 2σ from all DNA monomers, and short simulations with a soft repulsive potential were used to push off any initially overlapped proteins. Using the resulting configuration as an initial state, simulations were conducted at a given binding energy for 2×10^7 time steps. As before, after 10 time-step intervals, the distance d_{pr} of each protein from the DNA was determined and the result was tabulated as either a bound (1) or unbound (0) state according to whether d_{pr} was less than or greater than an ad hoc distance cut-off, chosen to be $d_{bd} = 1.5\sigma$.

Figure 4.4 shows a time-series plot of bound and unbound proteins for the case of $\epsilon = 1k_BT$, a relatively weak binding energy. As with the case of an extended DNA chain studied in the last chapter, the figure demonstrates that weakly-bound proteins spend the majority of time unbound to the DNA (executing 3-D diffusion) and make only intermittent visits to the DNA chain. These intermittent bound states are somewhat more clustered in nature than those observed for extended chains, perhaps due to transient caging of protein particles in locally crowded regions of the DNA coil. In this scenario, correlated rebinding of proteins may occur via hopping to proximal DNA chain elements before escape from the coil region. At high binding energy, sliding via curvilinear diffusion along the coiled DNA chain was the dominant mode of protein motion, as in the case of extended DNA chains studied in the last chapter. Figure 4.5 shows a time-series plot of bound and unbound proteins for the case of $\epsilon = 3k_BT$, a relatively strong binding energy. The figure clearly shows that these proteins spend the majority of time bound to the DNA (executing sliding diffusive motion), with occasional short-lived hops off the chain. Figures 4.4 and 4.5, together with Figures 3.9 and 3.8 from the previous chapter suggest that for the extremes of low and high binding energy, the qualitative behaviour for protein diffusion in the presence of extended and coiled DNA is very similar.

At intermediate values of binding energy, mixed 1-D and 3-D diffusion characterized by periods of sliding, hopping and three-dimensional diffusion is expected. Figure 4.6 shows a time-series plot of bound and unbound proteins for the case of $\epsilon = 2k_BT$. The

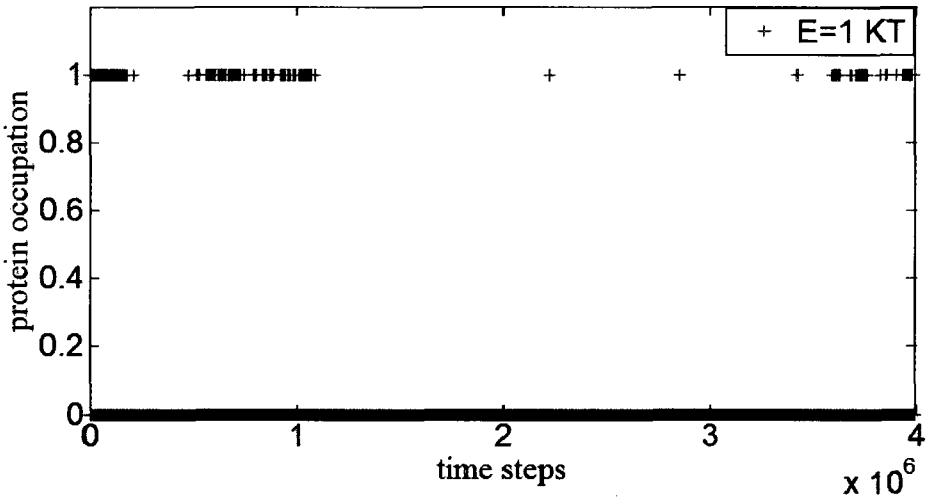


Figure 4.4: Protein occupation on a coiled DNA chain at different time-steps for $\epsilon = 1k_BT$.

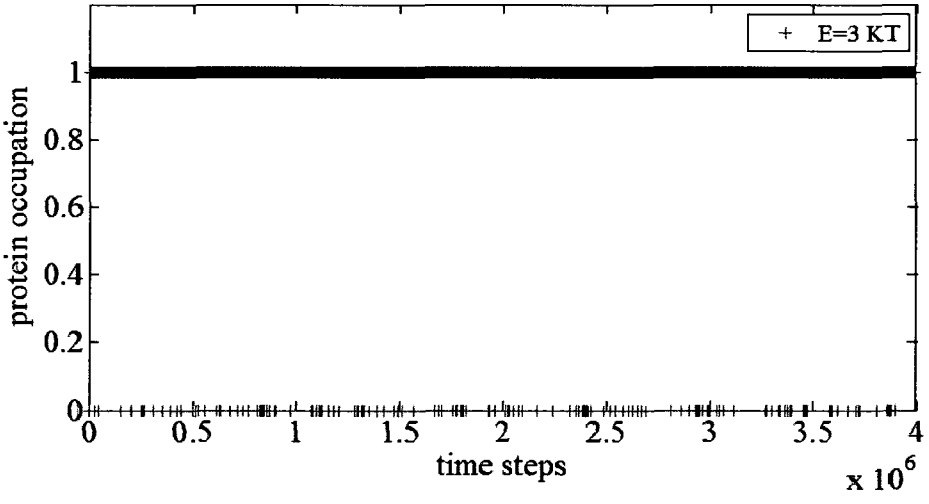


Figure 4.5: Protein occupation on the coiled DNA chain at different time-steps for $\epsilon = 3k_BT$.

figure indicates that these proteins spend both short and long periods on the DNA chain, supporting the hypothesis of mixed 1-D and 3-D diffusion punctuated by micro- and macro-hops. While this behaviour was also observed for extended DNA chains (see Figure 3.10 from the previous chapter), the coiled conformation of the DNA in this case may lead to inter-segmental transfer of proteins between sequence-distant DNA segments, a feature

that has previously been suggested to facilitate the target search process [28, 40, 41] and is supported by simulations described later in this chapter.

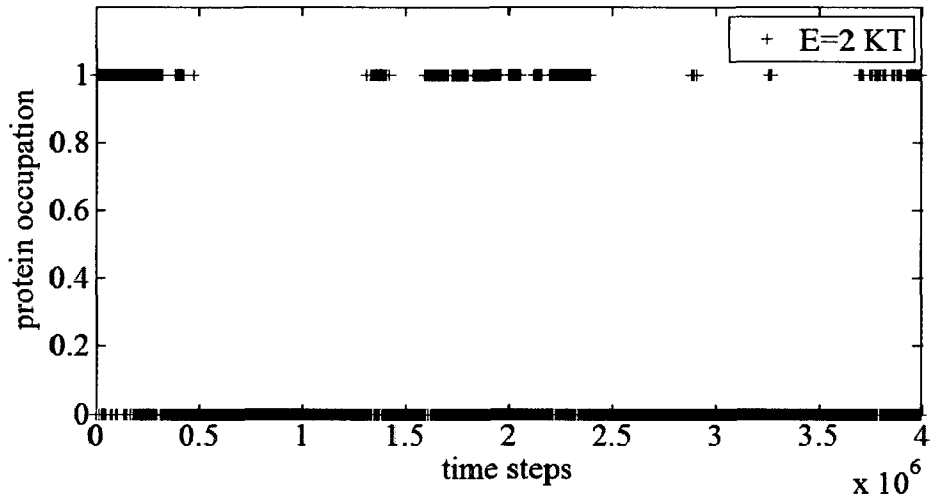


Figure 4.6: Protein occupation on the coiled DNA chain at different time-steps for $\epsilon = 2k_BT$.

Correlation of protein binding

Correlations between bound states at different time-steps provide information about the nature of the hopping process (e.g. short-lived hops(micro) vs macro-hops). Using the time-series data for bound and unbound protein states, the temporal autocorrelation between states was studied using the method described in the previous chapter. As before, the autocorrelation function $g(t)$ for bound states with different binding energies, defined by $g(t)=< on(t+t_0) \cdot on(t_0) >_{t_0}$, was calculated using times taken in increments of 10 time steps, giving $g(t)$ as an average over 2×10^6 events. The resulting un-normalized correlation between bound protein events is plotted versus time in Figure 4.7 for different binding energies, $\epsilon = 1, 2, 3k_BT$. Normalized $g(t)$ were then obtained by dividing the autocorrelation function by the ratio of the total number of time steps that protein is on the chain and the total number of protein states (time steps). Table 4.1 provides normalization data for the cases studied, and Figure 4.8 shows the normalized autocorrelation functions.

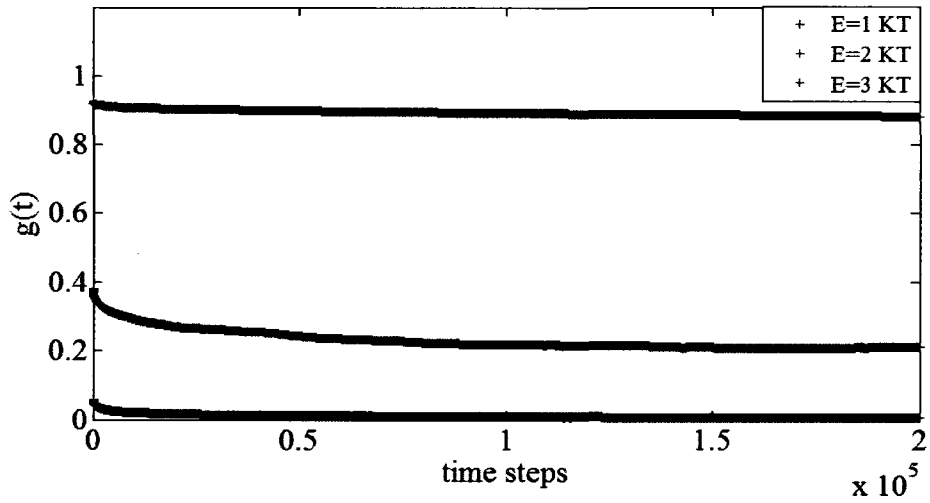


Figure 4.7: Correlation between on and off modes for proteins on coiled DNA with different binding energies, $\epsilon = 1, 2, 3k_BT$.

Binding Energy	$\epsilon = 1KT$	$\epsilon = 2KT$	$\epsilon = 3KT$
Total number of on states	115800	770422	1856200
Total number of states	2×10^6	2×10^6	2×10^6
Normalization factor	0.059	0.3852	0.9741

Table 4.1: $g(t)$ normalization data for three different binding energies.

Figure 4.8 shows that while bound states for high binding energies exhibit long-time correlation (indicative of 1-D sliding motion), as the binding energy decreases the bound states become more rapidly decorrelated in time, as expected for proteins that execute 3-D diffusive excursions off the coiled DNA chain. At the intermediate binding energy, $g(t)$ exhibits a rapid partial decay, followed by a long lived plateau after about 150000 time steps, presumably indicating the existence of inter-segmental transfer processes (which were not observed for the the case of extended DNA chains; As shown in Figure 3.12, where the decay occured continully).

The nature of these inter-segmental transfer processes will be clarified below in studies of the effect of DNA chain conformation on the target search time.

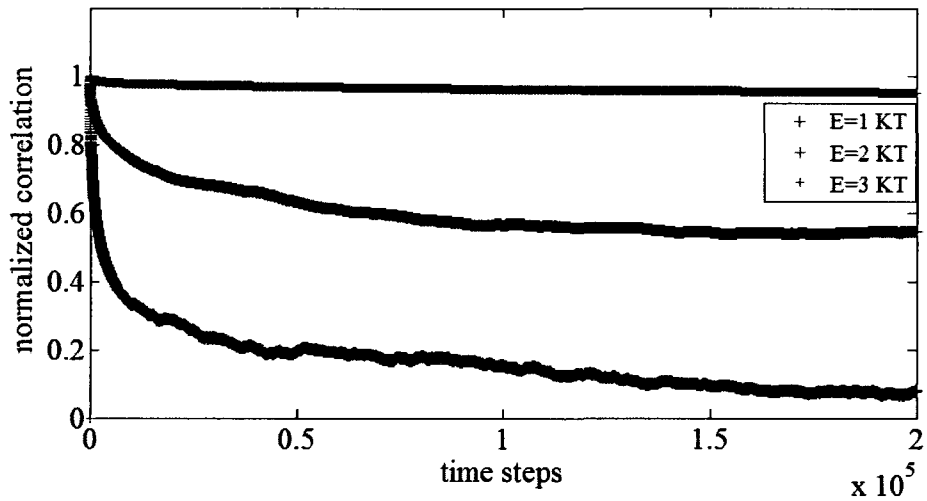


Figure 4.8: Normalized temporal autocorrelation function for proteins on coiled DNA with $\epsilon = 1, 2, 3k_BT$

4.2.2 Mean First Passage Time

Effects of DNA Conformation

The impact of DNA conformation on the protein target search process has not been studied in single molecule experiments, partly due to difficulties in spatial resolution of proteins in a coiled DNA chain. Simulation methods are therefore a potentially useful tool for investigating this issue. In this section we present a comparative study of protein target search for coiled and partially stretched DNA, with an emphasis on the mean first passage time (MFPT) and role inter-segmental jumps between two different segments of DNA plays in facilitated target localization. As before, a DNA chain of 500 monomers is chosen to allow for significant numbers of DNA loops in the relaxed coil state.

The coiled state is that shown in Figure 4.1, with the DNA centered in a simulation box of size $L_x \times L_y \times L_z = 50 \times 30 \times 32\sigma^3$ with periodic boundary conditions, and the target located in the middle of the coil (blue bead). Studies of the dependence of the MFPT on the binding energy ϵ were conducted for a collection of $N = 10$ randomly distributed proteins in a series of simulations, following the basic protocol of the previous chapter. The initial conditions for each run were prepared by randomly adding the proteins to the simulation

box and equilibrating the system in an initial run of 50000 time steps with repulsive interactions between protein molecules and DNA monomers. With the resulting configuration as an initial condition, each simulation was conducted with a particular attractive interaction between protein head monomers and DNA monomers ($\epsilon = 1k_BT$ to $4k_BT$) for a minimum of 6×10^6 time steps. This process was repeated 5000 times, each with a different initial condition. The time history of the protein coordinates was used to calculate the first passage time (FPT), the first instance of a protein finding the target DNA site. The mean first passage time (MFPT) was then obtained from the average over the runs of the FPT.

For comparison, analogous studies were conducted on partially extended DNA chains, which lack loop structures for inter-segmental transfer. The extended DNA chain was prepared from the coiled state shown in Figure 4.1 by increasing the end-to-end distance to 40% of its total contour length. This was achieved by applying a force of $f = 20\epsilon/\sigma$ to its end monomers during an equilibration simulation. The simulation box was also extended to $L_x \times L_y \times L_z = 200 \times 30 \times 32\sigma^3$, in order to accommodate the partially extended DNA chain. In order to match the *concentration* of proteins studied in the coiled configuration, 40 proteins were needed in the simulation box. Following the protocol using in the coiled case, MFPTs were obtained in the partially extended case for $\epsilon = 1k_BT$ to $4k_BT$ by averaging FPTs over the 5000 independent runs.

Figure 4.9 shows the dependence of the MFPT on binding energy ϵ for both coiled and extended DNA chains. In both cases, the MFPT exhibited a non-monotonic dependence on ϵ: as the binding energy increased, the MFPT first decreased, reaching an apparent minimum at an intermediate value around $3k_BT$, followed by a shallow increase (coiled case) or perhaps a plateau (extended case). This behaviour is qualitatively similar to the MFPT results of the previous chapter and the theoretical predictions of Refs [28, 40, 41] discussed previously (see Figure 3.16). Interestingly, while the MFPT values for the coiled and extended DNA conformations agree at low and high binding energy, the MFPTs for the coiled case are systematically smaller at intermediate binding energy. This behaviour is consistent with differences in the time-series data for bound and unbound states presented

earlier in this chapter. At low binding energies, we expect a regime of pure 3-D diffusion, independent of DNA chain conformation. At high binding energies, once bound proteins do not escape the DNA chain and are constrained to move by curvilinear diffusion (sliding) along the chain, which becomes progressively slower as the binding energy increases. Such sliding motion is also insensitive to the global chain conformation. At intermediate binding energies, we attribute the gap in MFPT between extended and coiled DNA to the effects of inter-segmental transfer that are operative in this regime. In the coiled state, such inter-segmental transfer between distant monomers in the sequence allows for a more efficient search of the DNA chain for its target, whereas inter-segmental transfer in the extended state is replaced by local hopping events.

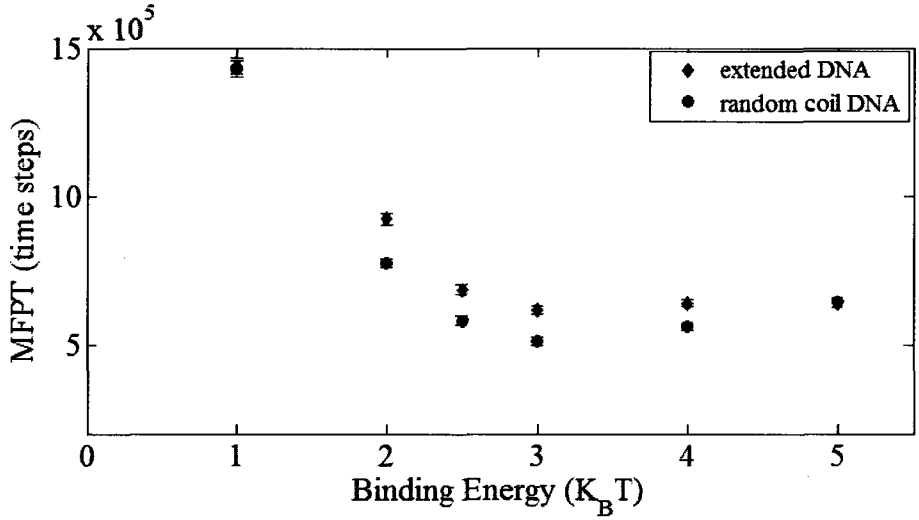


Figure 4.9: MFPT versus binding energy for proteins on coiled (black crosses) and extended (red stars) DNA chains. Error bars shown are the standard errors of the mean.

Inter-segmental transfer on a DNA loop

In this section, studies of the inter-segmental transfer of a protein molecule between two sites of DNA are presented. As discussed in the previous section, the binding energy between proteins and DNA is an important factor in controlling the inter-segmental transfer process in the case of coiled DNA. The frequency of inter-segmental transfer events is ex-

pected to depend on both the distance of two cognate sites and the binding strength of the protein with DNA. To investigate this dependence, a loop structure (comprised of 100 DNA monomers) was extracted from the DNA coil shown in Figure 4.1. Two DNA monomers in close proximity were chosen as the reference cognates, shown in blue in Figure 4.10.

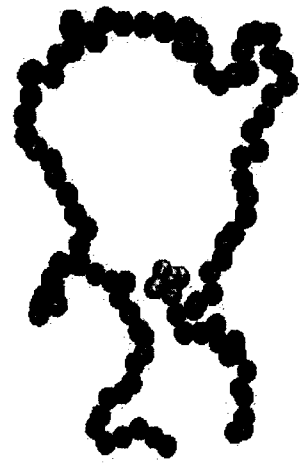


Figure 4.10: loop consisting of 100 monomers.two middle loops are shown by blue beads.

This loop was placed in a simulation box of size $L_x \times L_y \times L_z = 50 \times 25 \times 14\sigma^3$ with periodic boundary conditions. In order to investigate the effect of distance between cognate sites on the inter-segmental transfer process, a series of loop structures were prepared with blue cognate sites fixed in position with different separations from $r = \sigma$ to $r = 5\sigma$. Simulations were then performed on loops with a given cognate separation r in the presence of a single binding protein with $\epsilon = 1k_BT$ to $4k_BT$ as follows. The protein molecule was randomly placed within the simulation box and a long simulation was performed at fixed r and ϵ for 5×10^7 time-steps. The protein and DNA monomer coordinates were recorded and used to analyze the interaction of the protein with the DNA loop. The focus of the time-series analysis was on the events involving diffusive dynamics near the cognate sites. For each instance of the protein binding to one of the cognate sites, the elapsed time required

to bind to the other cognate site was tabulated. If this elapsed time is sufficiently short, the event may be identified as a direct inter-segmental hop. On the other hand, if this elapsed time is sufficiently long, the process is likely to be one involving either sliding diffusion along the loop (for large ϵ), large-scale 3-D diffusion (for small ϵ), or some combination of the two. In our analysis, we chose a cut-off time of $t_c = 300$ time steps, and assigned those events involving sequential visitation of the two cognate sites by the protein in a elapsed time less than t_c as inter-segmental transfers. The number of inter-segmental transfers that occurred during a simulation was used to calculate a hopping fraction, f, defined as the ratio of the number of times direct transfer between cognates occurs after a protein encounters one cognate to the number of times the protein encounters the initial cognate monomer. Figure 4.11 show a plot of the hopping fraction f as a function of the binding energy ϵ for two values of the cognate separation, $r = 3\sigma$ and 5σ.

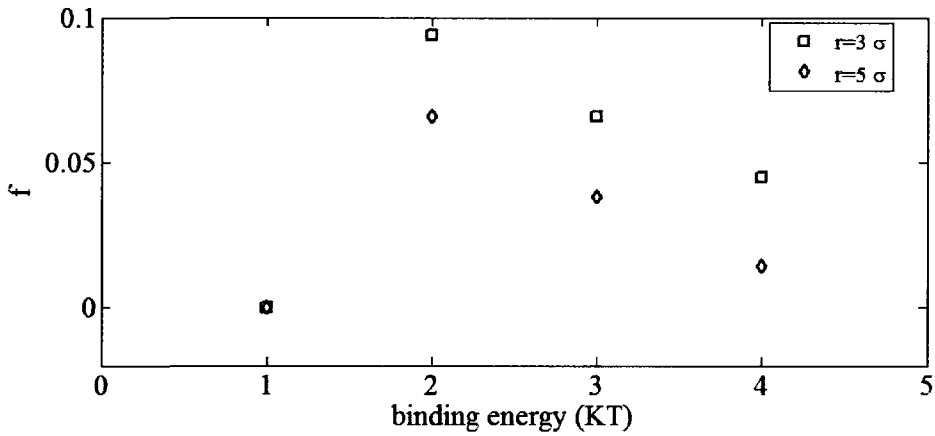


Figure 4.11: Inter-segmental transfer fraction of proteins between two cognate sites as a function of binding energy and the distance between the specific sites.

At the lowest binding energy, $\epsilon = 1k_BT$, f is essentially zero. Weakly binding proteins execute predominantly 3-D diffusion in the vicinity of the loop, rarely associating with the DNA, and thus precluding inter-segmental transfer events. This behaviour should be independent of the DNA cognate site separation, as verified by the simulation results. As the

binding energy increases, proteins eventually spend an increasing fraction of their time associated with the DNA and executing sliding diffusive motion and have fewer opportunities for performing hops off the DNA. Thus, as the binding energy increases from $\epsilon = 2k_BT$ to $\epsilon = 4k_BT$, the probability of inter-segmental transfer decreases, as reflected in the figure. As expected, proteins with intermediate binding energy have the highest f, as these conditions support the balance between sliding 1-D diffusion along the DNA chain and transient hops off the chain. The data in Figure 4.11 shows that the most efficient inter-segmental transfer occurs $\epsilon = 2k_BT$, in agreement with the results of Figure 4.9, in which the gap between MFPT for the coiled vs extended DNA chains was largest at $\epsilon = 2k_BT$.

4.3 Conclusion

In this chapter, the effects of protein conformation on DNA-facilitated protein transport and the target search process were investigated by Brownian dynamics simulation of a model coiled DNA chain in the presence of multi-particle protein molecules with selective binary interactions with DNA monomers. The loop structures present in this coiled DNA chain were shown to facilitate inter-segmental transfer of proteins between distant parts of the DNA sequence. Such inter-segmental transfer has been observed experimentally [34] and is thought to be one of the three main mechanisms governing the proteins target search process [11, 35].

At low and high values of the binding energy ϵ, the nature of protein dynamics in the presence of the DNA chain was remarkably similar for the cases of extended and coiled DNA. This similarity in behaviour was expected. For the case of low binding energy, the time series data showed that the proteins executed a 3-D diffusive search of the bulk that involved only transient interaction with the DNA chains. Thus, the conformation of the DNA in this limit was irrelevant. As a result there were no significant differences at low ϵ between the MFPT for target discovery on extended and coiled DNA chains. On the other hand, for high binding energy the time series data indicated that, after a protein first

binds to the DNA chain, it seldom unbinds but instead executes quasi 1-D sliding diffusion along the contour of the chain. This sliding process is only weakly dependent on the chain conformation, leading to very similar values of MFPT for coiled and extended DNA. One caveat here is the observation that the existence of loops with very narrow gaps can lead to more rapid sequence search due to contact-based transfer of proteins between distant monomers of the DNA sequence.

At intermediate binding energies, a clear difference in the behaviour of protein dynamics in the presence of the extended and coiled DNA chains was observed. At these energy scales, protein motion included a combination of 1-D sliding, bulk 3-D diffusion and small-to-medium scale hops between separated DNA monomers. The effect of these hops was different for extended and coiled DNA. In the former case, hops were local events involving modest jumps between monomers. Such events had relatively little impact on the search process. In the latter case, short-scale hops can occur across a loop gap resulting in an effectively long-range jump in the DNA monomer sequence. These so-called inter-segmental transfers were demonstrated explicitly in simulations of a DNA loop model, which showed that they may occur over a range of loop gaps.

The MFPT was found to be a non-monotonic function of binding energy for both extended and coiled DNA chains. At the lowest binding energy, $\epsilon = 1k_BT$, the search process was essentially a 3-D diffusive search for a single monomer target, a relatively slow process. With increasing binding energy, one-dimensional diffusion became increasingly dominant over three-dimensional diffusion, leading to a more efficient search as proteins became increasingly confined to the immediate vicinity of the DNA chain during the search, with occasional hops between close and distant monomers. However, once this binding energy was sufficiently large, the proteins became essentially confined to the chain. In this limit, hops and inter-segmental transfers were inhibited and the proteins spent more of their time oversampling small regions of DNA, thereby slowing the search process. Thus, an optimal binding energy, near $\epsilon = 3k_BT$, for minimizing the target search time was observed for both the cases of extended and coiled DNA, in qualitative agreement with theoretical expecta-

tions [28, 40, 41]. Moreover, in the intermediate binding energy range, $2k_BT \leq \epsilon \leq 3k_BT$, the MFPT was found to be systematically smaller for coiled DNA. As the primary difference in the protein dynamics between the two DNA conformations in this energy range is the emergence of inter-segmental transfers in the coiled DNA case, it seems likely that the inter-segmental transfer mechanism plays a role in facilitating the target search process, as suggested long ago in Ref. [10]. However, further experimental and theoretical studies are required to uncover the essential effects of inter-segmental transfer on the protein-DNA target search process.

Chapter 5

Discussion and Conclusion

In this thesis work, a relatively simple coarse-grained model of protein-DNA systems was utilized to investigate several essential features of DNA target searches that lead to facilitated diffusion. The proteins could sample the DNA contour through three fundamental search mechanisms including sliding, hopping/inter-segmental transfer, and 3-D diffusion. The interplay between the search mechanisms is controlled by the binding energy strength between the DNA and the proteins. For the case of extended DNA and single proteins, the most efficient search is when the sliding process dominates over two other mechanisms (3-D and hopping), whereas in the random coil DNA, the combination of 1-D sliding and 3-D motion leads to faster DNA sampling than either pure 1-D or 3-D motion. The MFPT dependency on binding energy was non-monatomic with an optimum value of about $\epsilon = 3K_BT$ for both the cases of extended and random coil DNA, in a qualitative agreement with theoretical results [40].

For the case of stretched DNA, the proposed model leads to a 1-D diffusion coefficient which was somewhat larger than typical experimental values. We obtained 5×10^{-7} cm^2/s, while the experimental mean value for the one-dimensional diffusion coefficient for a protein of size 10 nm is approximately D_{1d}=1.310^{-9} cm^2/s [33]. The discrepancy between these two values may be due to the fact that, in addition to binding interactions between DNA and proteins, real proteins can interact with the DNA through coordinated hydrogen bonds and electrostatic interactions when the protein is sufficiently close (but not actually

bound) to the DNA. The simulation model for a stretched DNA also leads to predictions different from those of other kinetic models. For instance, the MFPT required for a protein to find its specific cognate site on DNA by pure 1-D diffusion was found to decrease as a power law of the number of bound proteins $N^{-1.85}$, whereas a 1-D theoretical study [37] predicted a N^{-2} scaling. The difference may be due to the fact that in our model the proteins, which have repulsive interactions with other proteins, are embedded in 3-D and thus are able to pass each other by moving around the DNA chain axis, thereby softening the dependency of MFPT on protein concentration. Another computational model predicted non-monatonic behaviour of relative search time as a function of the protein the binding energy and concentration [40]. At low protein concentration, a protein could spend its time executing a rapid series of binding and un-binding events, thereby increasing the total search time, whereas our simulation shows that at low protein concentration, high binding energies make the protein slide along the DNA for a long period of time, facilitating the search process. A simulation approach such as ours also provides access to molecular level processes at a level of detail and resolution unavailable in current experimental approaches. For example, in a single molecule experiment [50], where the cumulative distribution of the jump lengths for different salt concentrations was investigated, only the translocations of the proteins larger than 200 nm were able to be visualized. Our simple model is able to measure very small jump lengths (about 0.05 σ) for a protein of size 2σ, which could more clearly specify the mechanisms involved in a search process.

Chapter 4 of the thesis treats the DNA chain as a coil fluctuating in three-dimensional space and interacting with multiple model protein molecules with tunable binding affinities. For random coil DNA, the DNA conformation or spatial organization has an important impact on facilitating the diffusion of proteins along a DNA coil, as found by previous experimental studies. At an intermediate binding energy $\epsilon = 2K_BT$, there is a balance between three fundamental search mechanisms that leads to inter-segmental transfers of proteins between two distant DNA segments; whereas at very high and low binding energy strengths, due to the low frequency of inter-segmental transfers, the DNA conformation

shows little effect on the search process, in accord with the experimental results in Ref. [35].

The standard deviation of the MFPT is always rather large, although the mean value itself did not change significantly. There is some theoretical justification for the inherent noise observed in our simulation data. Refs. [48, 49] argue that any quantity like the MFPT which depends on the random diffusion of a small collection of molecules in space, is inherently noisy. Even with the reduction of dimensionality characteristic of the sliding mechanism, the noisy behaviour of the MFPT is not noticeably reduced.

Although the results of the simulations of this model largely supported existing notions about the nature of the target search process and provided quantitative estimates not previously available in some cases, many questions still remain to be answered. This simple model provides no insight into how the protein recognizes and fixes to its specific target during 1-D sliding. Incorporating target recognition and specificity into the model will certainly require a finer and more realistic description of the DNA sequences and protein molecules. In future work we will focus on protein sliding along an extended double helical DNA chain with major and minor grooves, and re-investigate the MFPT as a function of protein binding affinity and concentration. Our current model is for DNA coils in free solution, for which all base pairs are uniformly accessible for protein binding (a case most relevant to Prokaryotes). However, native DNA in Eukaryotes is highly compacted into chromatin structures, in which portions of the DNA are already strongly bound to immobile protein complexes, making some regions of the DNA inaccessible for binding by other free proteins in solution. Therefore, it would be very interesting to model a compact DNA coil and investigate the effect of such initially bound proteins (obstacles) on the MFPT of soluble proteins [51].

References

[1] R.R. Sinden. *DNA structure and function.* Academic Pr, 1994.

[2] G. Orphanides and D. Reinberg. A unified theory of gene expression. *Cell*, 108(4):439–451, 2002.

[3] M. Ptashne and A. Gann. *Genes & signals.* CSHL Press, 2002.

[4] C.R. Calladine, H.R. Drew, and Travers A.A. *Understanding DNA: the molecule and how it works.* Academic Press, 2004.

[5] A.D. Bates and Maxwell A. *DNA TOPOLOGY.* Oxford Univ Press, 2005.

[6] J.D. Watson and Berry A. *The secret of life.* Arrow, 2004.

[7] A. Kornberg and T.A. Baker. *DNA replication.* WH Freeman New York, 1992.

[8] JM Schurr. The one-dimensional diffusion coefficient of proteins absorbed on DNA. Hydrodynamic considerations. *Biophysical chemistry*, 9(4):413, 1979.

[9] AD Riggs, S. Bourgeois, and M. Cohn. The lac repressor-operator interaction. 3. Kinetic studies. *Journal of molecular biology*, 53(3):401, 1970.

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

[11] PH Von Hippel and OG Berg. Facilitated target location in biological systems. *Journal of Biological Chemistry*, 264(2):675–678, 1989.

[12] C. Bai, C. Wang, X.S. Xie, and P.G. Wolynes. Single molecule physics and chemistry. *Proceedings of the National Academy of Sciences*, 96(20):11075, 1999.

[13] C. Bustamante, J.C. Macosko, and G.J.L. Wuite. Grabbing the cat by the tail: manipulating molecules one by one. *Nature Reviews Molecular Cell Biology*, 1(2):130–136, 2000.

[14] M.C. Williams. Optical tweezers: Measuring piconewton forces. *Biophysics Textbook Online: http://www. biophysics. org/btol*, 2002.

[15] A. Ashkin. Optical trapping and manipulation of neutral particles using lasers. *Proceedings of the National Academy of Sciences of the United States of America*, 94(10):4853, 1997.

[16] D.G. Grier. A revolution in optical manipulation. *Nature*, 424(6950):810–816, 2003.

[17] F. Ritort. Single-molecule experiments in biological physics: methods and applications. *Journal of Physics Condensed Matter*, 18(32):531, 2006.

[18] F. Brochard-Wyart. Deformations of one tethered chain in strong flows. *EPL (Europhysics Letters)*, 23:105–111, 1993.

[19] B. Ladoux and PS Doyle. Stretching tethered DNA chains in shear flow. *EPL (Europhysics Letters)*, 52:511–517, 2000.

[20] G.W. Slater, Y. Gratton, and M. Kenward. Deformation, stretching, and relaxation of single-polymer chains: fundamentals and examples. *Soft materials: structure and dynamics*, page 73, 2004.

[21] Y. Gratton and GW Slater. Molecular dynamics study of tethered polymers in shear flow. *The European Physical Journal E: Soft Matter and Biological Physics*, 17(4):455–465, 2005.

[22] D. Axelrod. Total internal reflection fluorescence microscopy in cell biology. *Traffic*, 2(11):764–774, 2001.

[23] GI Mashanov, D. Tacon, AE Knight, M. Peckham, and J.E. Molloy. Visualizing single molecules inside living cells using total internal reflection fluorescence microscopy. *Methods*, 29(2):142–152, 2003.

[24] D.M. Gowers, G.G. Wilson, and S.E. Halford. Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. *Proceedings of the National Academy of Sciences*, 102(44):15883–15888, 2005.

[25] P.C. Blainey, A.M. van Oijen, A. Banerjee, G.L. Verdine, and X.S. Xie. A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proceedings of the National Academy of Sciences*, 103(15):5752, 2006.

[26] J.H. Kim and R.G. Larson. Single-molecule analysis of 1D diffusion and transcription elongation of T7 RNA polymerase along individual stretched DNA molecules. *Nucleic Acids Research*, 2007.

[27] A. Granéli, C.C. Yeykal, R.B. Robertson, and E.C. Greene. Long-distance lateral diffusion of human Rad51 on double-stranded DNA. *Proceedings of the National Academy of Sciences*, 103(5):1221–1226, 2006.

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

[29] J. Gorman and E.C. Greene. Visualizing one-dimensional diffusion of proteins along DNA. *Nature Structural & Molecular Biology*, 15(8):768–774, 2008.

[30] S.E. Halford and J.F. Marko. How do site-specific DNA-binding proteins find their targets? *Nucleic acids research*, 32(10):3040, 2004.

[31] S. Halford. Hopping, jumping and looping by restriction enzymes. *Biochemical Society Transactions*, 29:363–374, 2001.

[32] K. Sakata-Sogawa and N. Shimamoto. RNA polymerase can track a DNA groove during promoter search. *Proceedings of the National Academy of Sciences*, 101(41):14731–14735, 2004.

[33] YM Wang, R.H. Austin, and E.C. Cox. Single molecule measurements of repressor protein 1D diffusion on DNA. *Physical review letters*, 97(4):48302, 2006.

[34] J. Iwahara and G.M. Clore. Direct observation of enhanced translocation of a homeodomain between DNA cognate sites by NMR exchange spectroscopy. *Journal of the American Chemical Society*, 128(2):404–405, 2006.

[35] B. van den Broek, MA Lomholt, S.M.J. Kalisch, R. Metzler, and GJL Wuite. How DNA coiling enhances target localization by proteins. *Proceedings of the National Academy of Sciences*, 105(41):15738, 2008.

[36] M.A. Lomholt, B. van den Broek, S.M.J. Kalisch, G.J.L. Wuite, and R. Metzler. Facilitated diffusion with DNA coiling. *Proceedings of the National Academy of Sciences*, 106(41):8204, 2009.

[37] I.M. Sokolov, R. Metzler, K. Pant, and M.C. Williams. Target search of N sliding proteins on a DNA. *Biophysical journal*, 89(2):895–902, 2005.

[38] K. Pant, R.L. Karpel, I. Rouzina, and M.C. Williams. Mechanical measurement of single-molecule binding rates: kinetics of DNA helix-destabilization by T4 gene 32 protein. *Journal of molecular biology*, 336(4):851–870, 2004.

[39] O. Givaty and Y. Levy. Protein sliding along DNA: dynamics and structural characterization. *Journal of Molecular Biology*, 385(4):1087–1097, 2009.

[40] AG Cherstvy, AB Kolomeisky, AA Kornyshev, et al. Protein- DNA Interactions: Reaching and Recognizing the Targets. *J. Phys. Chem. B*, 112(15):4741–4750, 2008.

[41] K.V. Klenin, H. Merlitz, J. Langowski, and C.X. Wu. Facilitated diffusion of DNA-binding proteins. *Physical review letters*, 96(1):18104, 2006.

[42] E. Dickinson and D.J. McClements. *Advances in food colloids*. Aspen Publishers, 1995.

[43] W. Coffey, Y.P. Kalmykov, and JT Waldron. *The Langevin equation: with applications to stochastic problems in physics, chemistry, and electrical engineering*. World Scientific Pub Co Inc, 2004.

[44] M. Kröger. *Models for polymeric and anisotropic liquids*. Springer Verlag, 2005.

[45] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[46] DC Rapaport. *The art of molecular dynamics simulation*. Cambridge Univ Pr, 2004.

[47] J. Gorman, A. Chowdhury, J.A. Surtees, J. Shimada, D.R. Reichman, E. Alani, and E.C. Greene. Dynamic basis for one-dimensional DNA scanning by the mismatch repair complex Msh2-Msh6. *Molecular cell*, 28(3):359–370, 2007.

[48] G. Tkačik and W. Bialek. Diffusion, dimensionality, and noise in transcriptional regulation. *Physical Review E*, 79(5):51901, 2009.

[49] R. Metzler. Keeping up with the noise. *Physics*, 2:36, May 2009.

[50] I. Bonnet, A. Biebricher, P.L. Porte, C. Loverdo, O. Benichou, R. Voituriez, C. Escude, W. Wende, A. Pingoud, and P. Desbiolles. Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA. *Nucleic Acids Research*, 36(12):4118, 2008.

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

Appendix A

```
Lammps input script for proteins as single beads
#3d simulation
# General parameters for the system
units           lj
atom_style      bond
boundary        p p f
dimension       3
# Neighbor list parameters
neighbor        0.3 bin
neigh_modify    delay 1
#####################################################################
# Selected parameters
#####################################################################
# Loop over index of data file on multiple partitions
# Using a universe variable
variable      i universe 1 2 3 4 5 6 7 8 9 10  11 12 13 14 15 16
 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58
 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
# Reading data file
read_data       data-$i
# Group of atoms
group           chain type 1
group           end1 type 2
group           end2 type 3
group           protein type 4
# Bond potential parameters
bond_style      fene
bond_coeff      1 100.0 1.2 1.0 1.0
# to push off the overlapped particles
pair_style      soft 1.5
pair_coeff      * * 1.0 30.0 1.5
# Timestep
timestep       0.005
#Create initial velocities
velocity        all create 1.0 2349852
```

```
# Fix both ends of DNA
velocity       end1 set 0.0 0.0 0.0 units box
velocity       end2 set 0.0 0.0 0.0 units box
fix            2 end1 setforce 0.0 0.0 0.0
fix            3 end2 setforce 0.0 0.0 0.0
# Run the simulation in the NVT ensemble
fix            1 all nve
fix            4 all langevin 1.0 1.0 0.1  135786
# reflective walls in z direction
fix            5 all wall/reflect zlo zhi
# Run a few steps to prepare the system
run           5000
# Pairwise potential parameters
pair_style    lj/cut 3.75
pair_coeff    * * 1.0 1.0 1.12246
pair_coeff    1 4 3 1.5 3.75
pair_coeff    2 4 3 1.5 3.75
pair_coeff    3 4 3 1.5 3.75
pair_coeff    4 4 4.0 2.0 2.22448
# Dump system config every 50 steps
dump           1 protein xyz 50 protein-$i.xyz
dump           2 all dcd 10000 all-$i.dcd
dump           3 chain xyz 50 chain-$i.xyz
dump           4 end1 xyz 50 end1-$i.xyz
# Run simulation
run           5000000
#End of loop over data files
clear
next                    i
jump              /home/nazanin/Documents/mean-bead

Lammps input script for proteins medeled as shells
#3d simulation
# General parameters for the system
units          lj
atom_style     bond
dimension      3
special_bonds   0 1 1
################################################################
# Selected parameters
################################################################
# Loop over index of data file on multiple partitions
# Using a universe variable
variable     i universe 1 2 3 4 5 6 7 8 9 10  11 12 13 14 15 16
 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58
 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
# Reading data file
```

```
read_data      data-random-coil-$i
# Group of atoms
group          chain type 1
group          end1 type 2
group          end2 type 3
group          head type 5
group          target type 6
group          tail type 7
# Bond potential parameters
bond_style     hybrid fene harmonic
bond_coeff     1  fene 100.0 1.2 1.0 1.0
bond_coeff     2  harmonic 500 0.92
bond_coeff     3  harmonic 500 0.92
# Pairwise potential parameters
pair_style     lj/cut  1.12254
pair_coeff     * * 1.0 1.0 1.12246
# Timestep
timestep      0.005
#Create initial velocity
velocity       all create 1.0 2349852
# Fix both ends of DNA
velocity       end1 set 0.0 0.0 0.0 units box
velocity       end2 set 0.0 0.0 0.0 units box
fix            2 end1 setforce 0.0 0.0 0.0
fix            3 end2 setforce 0.0 0.0 0.0
# Run the simulation in the NVT ensemble
fix            1 all nve
fix            4 all langevin 1.0 1.0 0.1  135786
# Run a few steps to prepare the system
run            50000
# Pairwise potential parameters
pair_style     lj/cut  3.0
pair_coeff     * * 1.0 1.0 1.12246
pair_coeff     1 5 4.0 1.0 3.0
pair_coeff     2 5 4.0 1.0 3.0
pair_coeff     3 5 4.0 1.0 3.0
pair_coeff     5 6 4.0 1.0 3.0
# Dump system config every 50 steps
dump           1 shell xyz 50 head-$i.xyz
dump           2 end1 xyz 50 target-$i.xyz
dump           2 end1 xyz 50 chain-$i.xyz
dump           3 all dcd 50 random-coil-$i.dcd
# Run simulation
run          5000000
#End of loop over data files
clear
next                    i
jump              /home/nazanin/Documents/mean-random-coil
```

Appendix B

```
A sample of the initialization programs (create-data)
#!/usr/bin/perl -w

$xstart = -240;
$ystart=8;
$zstart=16;
$minR = 4.5;
for ($i=0;$i<10;$i++) {
        $xpos[$i] = $xstart + rand()*40;
        $ypos[$i] = $ystart + rand()*20;
        $zpos[$i] = $zstart ;
        for ($j=0;$j<$i;$j++) {
              if(sqrt(($xpos[$i]-$xpos[$j])*($xpos[$i]-$xpos[$j])+
              ($ypos[$i]-$ypos[$j])*($ypos[$i]-$ypos[$j])+
              ($zpos[$i]-$zpos[$j])*($zpos[$i]-$zpos[$j]))<$ minR)
       {
                        $i--;
                        $j = $i;
       }
                                }
$x[$i]=$xpos[$i]+1;
$y[$i]=$ypos[$i]+1;
$z[$i]=$zpos[$i]+1;
$xl[$i]=$xpos[$i]-1;
$yl[$i]=$ypos[$i]-1;
                    }
print "
560 atoms
629 bonds
6 atom types
3 bond types

-245  -195 xlo xhi
1  30  ylo yhi
-12  20  zlo zhi

Masses
```

```
1 1.0
2 1.0
3 1.0
4 1.0
5 1.0
6 1.0
#Chain coordinates
Atoms
1 1 2 -211.269 19.2102 -9.82276
2 1 1 -211.369 19.2744 -8.9644
3 1 1 -210.497 19.1437 -9.0161
4 1 1 -209.799 18.6624 -9.33923
5 1 1 -209.012 18.5823 -8.91988
6 1 1 -209.312 17.8615 -8.4575
7 1 1 -208.76 17.4428 -7.95438
8 1 1 -208.374 17.4159 -7.17025
9 1 1 -208.17 18.2852 -7.24164
10 1 1 -207.857 19.087 -6.93513
.
.
.
250 1 6 -220.193 17.1144 7.32854
251 1 1 -219.8 16.3235 7.34137
252 1 1 -219.485 15.6077 7.73344
253 1 1 -218.889 15.964 8.28711
254 1 1 -218.196 16.3596 8.695
255 1 1 -218.678 16.0718 9.37395
256 1 1 -218.314 15.4968 9.91855
257 1 1 -218.834 15.8655 10.5176
258 1 1 -219.261 16.1068 11.2234
259 1 1 -219.106 15.4004 11.7171
260 1 1 -219.832 14.9145 11.8681
:
:
:
480 1 1 -228.707 26.0597 -10.8105
481 1 1 -228.267 25.3506 -10.5194
482 1 1 -227.845 25.1701 -11.2411
483 1 1 -227.436 25.9492 -11.4235
484 1 1 -226.718 25.9987 -11.9075
485 1 1 -226.017 26.5179 -11.6932
486 1 1 -226.436 27.267 -11.5062
487 1 1 -226.065 27.9494 -11.1742
488 1 1 -225.973 27.7295 -10.3215
489 1 1 -226.356 27.0416 -9.93684
490 1 1 -225.79 26.429 -9.69492
491 1 1 -226.328 25.7088 -9.81134
492 1 1 -226.206 24.8489 -9.67426
493 1 1 -226.222 24.0997 -9.22107
```

```
494 1 1 -225.668 24.2585 -8.54556
495 1 1 -225.373 25.0788 -8.40131
496 1 1 -225.02 25.7586 -7.89692
497 1 1 -225.348 26.4119 -8.42099
498 1 1 -225.605 27.191 -8.14919
499 1 1 -224.878 27.5889 -7.9027
500 1 3 -225.065 28.2293 -7.32667
# Proteins coordinates
501     1       5       $xpos[0]        $ypos[0]        $zpos[0]
502     1       4       $x[0]           $ypos[0]        $zpos[0]
503     4       4       $xpos[0]        $y[0]           $zpos[0]
504     4       4       $x1[0]          $ypos[0]        $zpos[0]
505     4       4       $xpos[0]        $y1[0]          $zpos[0]
506     4       4       $xpos[0]        $ypos[0]        $z[0]
507     4       5       $xpos[1]        $ypos[1]        $zpos[1]
508     4       4       $x[1]           $ypos[1]        $zpos[1]
509     4       4       $xpos[1]        $y[1]           $zpos[1]
510     4       4       $x1[1]          $ypos[1]        $zpos[1]
511     4       4       $xpos[1]        $y1[1]          $zpos[1]
512     4       4       $xpos[1]        $ypos[1]        $z[1]
.
.
.
555     4       5       $xpos[9]        $ypos[9]        $zpos[9]
556     4       4       $x[9]           $ypos[9]        $zpos[9]
557     4       4       $xpos[9]        $y[9]           $zpos[9]
558     4       4       $x1[9]          $ypos[9]        $zpos[9]
559     4       4       $xpos[9]        $y1[9]          $zpos[9]
560     4       4       $xpos[9]        $ypos[9]        $z[9]
Bonds
1 1 1 2
2 1 2 3
3 1 3 4
4 1 4 5
5 1 5 6
6 1 6 7
7 1 7 8
8 1 8 9
9 1 9 10
10 1 10 11
11 1 11 12
12 1 12 13
13 1 13 14
14 1 14 15
15 1 15 16
16 1 16 17
17 1 17 18
18 1 18 19
19 1 19 20
```

```
.
.
.
481 1 481 482
482 1 482 483
483 1 483 484
484 1 484 485
485 1 485 486
486 1 486 487
487 1 487 488
488 1 488 489
489 1 489 490
490 1 490 491
491 1 491 492
492 1 492 493
493 1 493 494
494 1 494 495
495 1 495 496
496 1 496 497
497 1 497 498
498 1 498 499
499 1 499 500
# Head-Tail bonds
500 2 501 502
501 2 501 503
502 2 501 504
503 2 501 505
504 2 502 503
505 2 503 504
506 2 504 505
507 2 505 502
508 3 501 506
509 2 502 506
510 2 503 506
511 2 504 506
512 2 505 506
.
.
.
617 2 555 556
618 2 555 557
619 2 555 558
620 2 555 559
621 2 556 557
622 2 557 558
623 2 558 559
624 2 559 556
625 3 555 560
626 2 556 560
```

```
627 2 557 560
628 2 558 560
629 2 559 560
";
exit;
```

Appendix-C

to find proteins positions relative to DNA

```
#include <string.h>
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#include <time.h>
FILE *fp, *fq, *fz;
int main(){
float  pos, x,w1, y,z,x1,y1,z1,pos1;
char   str[10],str1[10];
int    m=0,n=2,i,j,b,k,w,c,d,f,r=3,a=500,on=0,off=0;
float x_pos[a], y_pos[a],x1_pos[r],
y1_pos[r],z_pos[a],z1_pos[r],w1_pos[r],w_pos[a];

            if ((fp = fopen("chain.xyz","r")) != NULL )
                  if ((fq = fopen("head.xyz","r")) != NULL )

             for(k=1;k<=2000000;k++){
m=0;
# read head coordinates from head.xyz

 for(j=1;j<=1;j++)
          {

          fscanf(fq, "%f",&pos1);
          fscanf(fq, "%s",&str1);
          fscanf(fq, "%f", &w1);
          w1_pos[j]= w1;
          fscanf(fq, "%f", &x1);
          x1_pos[j]= x1;
          fscanf(fq, "%f/n", &y1);
          y1_pos[j]= y1;
          fscanf(fq, "%f/n", &z1);
          z1_pos[j]= z1;
          }
# read chain coordinates from chain.xyz
```

```
 fscanf(fp, "%f",&pos);
 fscanf(fp, "%s",&str);
 for(i=1;i<=a;i++)
          {
          fscanf(fp, "%f",&w);
          w_pos[i]= w;
          fscanf(fp, "%f", &x);
          x_pos[i]= x;
          fscanf(fp, "%f/n", &y);
          y_pos[i]= y;
          fscanf(fp, "%f/n", &z);
          z_pos[i]= z;
          }
#open a file relative.xyz to append the data

 if ((fz = fopen("relative.xyz", "a")) != NULL ){
           for(i=1;i<=a;i++){
             if(sqrt((x_pos[i]-x1_pos[1])*(x_pos[i]-x1_pos[1])+
             (y_pos[i]-y1_pos[1])*(y_pos[i]-y1_pos[1])+
             (z_pos[i]-z1_pos[1])*(z_pos[i]-z1_pos[1]))<=1.5 )
             {
# print that protein is on the chain

                  fprintf(fz,"1\n");
                   on++;
                   m=1;
                   break;
             }
                           }
if(m!=1){
# protein is off the chain

fprintf(fz,"0\n");
       }
}
fclose(fz);
}
printf("%i\n",on);
fclose(fq);
fclose(fp);
}

Calculate mean square displacement of a protein (MSD)

#include <string.h>
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
```

```
#include <time.h>
#include <unistd.h>
FILE   *fq,*fp;
 main(){
int j,i,k;
double *x,*p,*w,*y,*z,*f;
float m,sum,w1,x1,y1,z1,f1,frame;
# read the unwraped protein coordinates from unwrap-head.xyz

 if ((fq = fopen("unwrap-head.xyz","r")) != NULL )
x=(double*)calloc(10000000,sizeof(double));
p=(double*)calloc(10000000,sizeof(double));
y=(double*)calloc(10000000,sizeof(double));
z=(double*)calloc(10000000,sizeof(double));
w=(double*)calloc(10000000,sizeof(double));
f=(double*)calloc(10000000,sizeof(double));

    for(i=1;i<=1500000;i++)
          {
         fscanf(fq, "%f", &w1);
          w[i]= w1;
          fscanf(fq, "%f", &x1);
          x[i]= x1;
          fscanf(fq, "%f/n", &y1);
          y[i]= y1;
          fscanf(fq, "%f/n", &z1);
          z[i]= z1;
          fscanf(fq, "%f/n", &f1);
          f[i]= f1;
          }
p[1]=0;
k=1;
 for(j=1;j<1500000;++j)
     {
    for(i=1;i<1500000;++i)
      {
      if (j+i=1500000){
  break;
                }
m=(x[i+j]-x[i])*(x[i+j]-x[i])+(y[i+j]-y[i])*(y[i+j]-y[i])+
        (z[i+j]-z[i])*(z[i+j]-z[i]);
  p[j]=p[j]+m;
      }
# append mean square displacement to msd-head.xyz
    if((fp=fopen("msd-head.xyz","a"))!=NULL)
    fprintf(fp,"%f  %f  \n",j , p[(j)]/(i-1));
    fclose(fp);
    }
}
```

Appendix-D

```
for((j=1;j<=50000;j++)){
#loop over 120 independant simulation
for((i=1;i<=120;i++)){
#create initial data (Appendix B)
./create-data >> data-random-coil-$i
mkdir 110-3-$j
#run 120 lammps jobs
mpirun -np 120 /home/nazanin/Documents/lmp_mvapich<mean-random-coil>
out-bead -partition 120x1 -in /home/nazanin/Documents/mean-random-coil
for((i=1;i<=120;i++)){
mv chain-$i.xyz 110-3-$j
mv head-$i.xyz 110-3-$j
mv target-$i.xyz 110-3-$j
rm data-$i
}
}
Calculate mean first passage time (MFPT.c)

#include <string.h>
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#include <time.h>
FILE *fp, *fq, *fz;
int main(){
float pos, x,w1, y,z,x1,y1,z1,pos1;
char   str[10],str1[10];
int    n=2,i,j,b,k,w,c,d,f;
float x_pos[a], y_pos[a],x1_pos[r], y1_pos[r],z_pos[a]
      ,z1_pos[r],w1_pos[r],w_pos[a];
#open files bead.xyz and target.xyz to read
         if ((fp = fopen("Head.xyz","r")) != NULL )
            if ((fq = fopen("target.xyz","r")) != NULL ){

             for(k=1;k<=50000;k++){
#read target coordinates in each time step
           for(j=1;j<=1;j++)
```

```
          {
          fscanf(fq, "%f",&pos1);
          fscanf(fq, "%s",&str1);
          fscanf(fq, "%f", &w1);
          w1_pos[j]= w1;
          fscanf(fq, "%f", &x1);
          x1_pos[j]= x1;
          fscanf(fq, "%f/n", &y1);
          y1_pos[j]= y1;
          fscanf(fq, "%f/n", &z1);
          z1_pos[j]= z1;
          }
#read all of the proteins coordinates in each time step
          fscanf(fp, "%f",&pos);
          fscanf(fp, "%s",&str);
          for(i=1;i<=a;i++)
          {
          fscanf(fp, "%f",&w);
          w_pos[i]= w;
          fscanf(fp, "%f", &x);
          x_pos[i]= x;
          fscanf(fp, "%f/n", &y);
          y_pos[i]= y;
          fscanf(fp, "%f/n", &z);
          z_pos[i]= z;
          }
# open a file MFPT.xyz to append the data

 if ((fz = fopen("MFPT.xyz", "a")) != NULL ){
          for(i=1;i<=a;i++){
               if(sqrt((y_pos[i]-y1_pos[1])*(y_pos[i]-y1_pos[1])+
               (x_pos[i]-x1_pos[1])*(x_pos[i]-x1_pos[1])+
               (z_pos[i]-z1_pos[1])*(z_pos[i]-z1_pos[1]))<=1.5)
                       {
                        fprintf(fz,"%i\n",k);
                        break;
                       }
                          }
                                          }
fclose(fz);
}
}
fclose(fq);
fclose(fp);
}
#!/bin/sh
for((i=1;i<=120;i++)){
mv head-$i.xyz head.xyz
mv target-$i.xyz target.xyz
```

```
#compile C file called MFPT.C
cc -g -lm  MFPT.c -o MFPT
./MFPT
#write the output file in MFPT.xyz
mv MFPT.xyz MFPT-$i
head -1 MFPT-$i >> MFPT-3KT.xyz
rm MFPT-$i
}
```