Modeling DNA and DNA-protein interactions

Ana Maria Florescu

UNIVERSITÉ DE GRENOBLE

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE GRENOBLE
Spécialité **Physique /Physique pour les sciences du vivant**
Arrêté ministériel : 7 août 2006

Présentée et soutenue publiquement par

Ana-Maria FLORESCU

le **22 Novembre 2010**

MODELISATION DE L'ADN ET DES INTERACTIONS ADN-PROTEINES

Thèse dirigée par **Marc JOYEUX**

JURY

Maria BARBI	Maître de Conférences Paris VI	Examinateur
Ralf BLOSSEY	Directeur de Recherche CNRS	Rapporteur
Marc Joyeux	Chargé de Recherche CNRS	Examinateur
Thierry DOMBRE	Professeur UJF Grenoble	Président
Michel PEYRARD	Professeur ENS Lyon	Rapporteur

Thèse préparée au sein du Laboratoire de Spectrométrie Physique
dans l' Ecole Doctorale de Physique de Grenoble

Acknowledgements

I would like to thank my supervisor, Marc Joyeux, for his encouragements, his patience and his constant availability. During these three years he has thought me how to handle a research subject, how to construct a model and how to lucidly analyze the results I obtain with it. He has also been a French language teacher, a running coach and a friend when needed. I was lucky to have such an advisor.

I would also like to thank Maria Barbi, Ralf Blossey, Thierry Dombre and Michel Peyrard for taking the time to read and evaluate this work and for their comments and suggestions regarding the manuscript.

I am grateful to the people from Laboratoire de Spectrométrie Physique for their help and their availability, and to the students for all the coffee and lunch breaks we shared. A special mention goes to Teo, for the same reasons and for always being here whenever I needed someone to talk to.

Finally, I would like to thank my parents, for always encouraging me to believe in myself.

Résumé

L'ADN est une des molécules qui est le plus étudiée, que ce soit en biologie, en chimie ou en physique, parce qu'elle code dans la séquence de nucléotides qui la compose l'information nécessaire à la synthèse des protéines. Les développements récents des expériences de microscopie ont permis l'étude d'un grand nombre de processus cellulaires, dont les plus connus sont la transcription et la réplication de l'ADN. Ces processus sont contrôlés par des protéines appelées *facteurs de transcription*, qui se lient à la séquence ADN et séparent ensuite localement les deux brins de la structure hélicoïdale pour accéder à l'information génétique. La dénaturation de l'ADN, c'est à dire la séparation des deux brins, est en elle-même un processus très intéressant pour la physique statistique, puisqu'elle peut être assimilée à une transition de phase.

Ma thèse est structurée en deux parties : la première partie (chapitres 2 à 4) porte sur la modélisation de la dénaturation de l'ADN, alors que la seconde partie (chapitres 5 à 8) propose et discute un modèle pour les interactions entre l'ADN et les protéines visant à décrire comment certaines protéines, comme les facteurs de transcription, trouvent leur cible dans la séquence d'ADN.

Après une introduction générale (chapitre 1), le deuxième chapitre de la thèse est une introduction portant plus particulièrement sur la dénaturation de l'ADN. Je commence par un rappel de la structure chimique et de la fonction de l'ADN, avant de décrire plusieurs modèles qui ont été développés pour l'étude de sa dénaturation. Je présente tout d'abord le modèle de Poland-Scheraga, qui fait partie de la catégorie des modèles statistiques (pour lesquels une paire de bases est décrite par une variable à deux états : ouvert ou fermé), puis deux modèles dynamiques, basés sur des expressions explicites du Hamiltonien du système en fonction des coordonnées et des vitesses des particules qui composent la séquence.

Le troisième chapitre décrit un modèle des expériences d'électrophorèse en deux dimensions. L'électrophorèse est une technique de séparation des séquences ADN qui est comparativement peu coûteuse et fréquemment utilisée en biologie. Elle est basée sur le fait que, dans un gel soumis à un champ électrique, les molécules d'ADN migrent avec des vitesses différentes en fonction de leur longueur et de leur composition. Cette séparation se fait généralement en deux étapes: tout d'abord une séparation en fonction de la longueur de la

séquence, puis une séparation en fonction de sa composition, provoquée par la dénaturation chimique ou thermique des molécules d'ADN. Le modèle discuté ici a été construit autour de MeltSim, un logiciel gratuit de calcul des courbes de dénaturation thermique de l'ADN basé sur le modèle de Poland-Scheraga. L'ajustement des paramètres du modèle a permis de prédire des positions des fragments d'ADN à la fin de la séparation en très bon accord avec les valeurs mesurées.

Dans le chapitre 4, je décris enfin comment j'ai pu améliorer un modèle dynamique de la dénaturation de l'ADN, qui avait été développé dans notre groupe avant le début de ma thèse. J'ai obtenu un nouveau jeu de paramètres pour ce modèle, qui permet de reproduire correctement plus de données expérimentales que précédemment, comme par exemple la force critique lors d'expériences de dénaturation mécanique, ou encore l'évolution de la température critique en fonction de la taille des séquences. J'ai aussi comparé les résultats obtenus grâce à ce modèle avec ceux obtenus à partir des modèles statistiques et je me suis finalement intéressé à l'ordre de la transition de phase de dénaturation prédit par ce modèle.

La deuxième partie du manuscrit porte sur la modélisation des interactions entre l'ADN et les protéines et sur les procédés par lesquels les protéines recherchent leur cible dans la séquence d'ADN. C'est là l'un des problèmes les plus discutés de la biophysique actuelle. En général, ces protéines trouvent très rapidement sur leur cible. Il est souvent admis que la méthode de recherche utilisée par les protéines, à savoir une combinaison de glissement 1D le long de l'ADN et de diffusion 3D dans la cellule, est beaucoup plus rapide que la diffusion 3D normale. Cependant les expériences montrent que cela est dû à la présence d'interactions électrostatiques entre la protéine et l'ADN et que la vitesse de recherche est par conséquent très fortement corrélée à la salinité du solvant.

Je commence, au chapitre 5, par une présentation du problème et de certains résultats expérimentaux, ainsi que par une courte discussion des modèles existants. Je propose également une courte introduction théorique à la recherche par des marches aléatoires.

Dans le sixième chapitre du rapport, je développe le modèle dynamique que je propose pour la description des interactions entre l'ADN et les protéines. L'ADN est décrite par un modèle "beads and springs" emprunté à la physique des polymères, alors que la protéine est modélisée par une sphère rigide dotée d'une charge ponctuelle en son centre. Le potentiel d'interaction ADN-protéine est composé de deux termes : un terme attractif d'origine

électrostatique et un terme répulsif de volume exclu, dont la somme présente un minimum à une distance de l'axe de l'ADN égale à la somme des rayons des beads décrivant l'ADN et la protéine. J'ai étudié les propriétés de ce modèle en intégrant les équations d'évolution du mouvement grâce à un algorithme de dynamique brownienne. Je présente les résultats concernant les mouvements 1D et 3D de la protéine obtenus avec ce model et je discute leur accord avec les expériences. Ce modèle prédit une accélération maximale de la recherche due à la diffusion facilitée de l'ordre de deux, c'est à dire nettement plus faible que celle prédite par certains modèles cinétiques.

Le chapitre 7 propose une amélioration de la description de la protéine comme un ensemble interconnecté de 13 sphères plutôt qu'une sphère unique. J'ai utilisé ce modèle pour vérifier les résultats présentés dans le chapitre précédent, mais également pour étudier l'influence des propriétés de la protéine sur la diffusion facilitée. Les simulations conduites avec ce nouveau modèle confirment la faiblesse de l'augmentation de la vitesse de recherche due à la diffusion facilitée par rapport aux prédictions de certains modèles cinétiques. Ce modèle montre également que la forme et l'élasticité de la protéine semblent n'affecter la vitesse du processus de recherche que par le biais de leur effet sur les valeurs du coefficient de diffusion. Enfin, ce modèle prédit que l'efficacité de la diffusion facilitée est influencée plutôt par la charge totale de la protéine que par les charges partielles placées sur les différents sites et que le glissement 1D de la protéine le long de l'ADN est souvent sous-diffusif.

Dans le dernier chapitre, je compare enfin les résultats obtenus grâce à mon modèle aux prédictions d'un des rares modèles cinétiques vraiment prédictif et je montre que les deux types de modèles s'accordent en fait pour prédire que la diffusion facilitée n'accélère pas toujours l'association entre l'ADN et les protéines. En fait, je montre même que la combinaison de glissement 1D le long de l'ADN et de diffusion 3D dans la cellule ne peut être plus efficace que la diffusion 3D normale que si le coefficient de diffusion 1D est plus grand que le coefficient de diffusion 3D, alors qu'on sait expérimentalement qu'il est entre 3 et 5 ordres de grandeur plus petit. Ces résultats sont en bon accord avec une relecture récente des résultats expérimentaux.

En conclusion, dans la première partie de mon travail j'ai utilisé le modèle de Poland-Scheraga pour décrire la séparation des séquences ADN par électrophorèse en deux dimensions, puis j'ai obtenu un meilleur jeu de paramètres pour le modèle dynamique de dénaturation de l'ADN développé dans notre groupe. Dans la deuxième partie, j'ai proposé un modèle dynamique pour l'étude des interactions entre l'ADN et les protéines. J'ai montré que ce modèle présente de

la diffusion facilitée et qu’il prédit un mouvement de la protéine en globalement bon accord avec les résultats expérimentaux. Cependant, les modèles décrivant la dénaturation ne peuvent pas être utilisés pour décrire l'interaction entre l'ADN et les protéines, et vice versa. La perspective essentielle de ce travail consiste donc à établir des modèles d'ADN et de protéines plus résolus, capables de décrire les deux phénomènes à la fois. Ce type de modèle autorisera également l'étude des interactions "spécifiques", c'est à dire dépendant de la séquence, qui permettent à la protéine de se fixer sur sa cible.

Table of contents

Part II: DNA-protein interactions

1. Introduction

One of the aims of computer simulations in science is the study of the properties of molecules and the interactions between them. They help understanding experimental results and sometimes even complement them. With the increasing development of computers, it is now possible to turn to the study of large systems, like bulk fluids or polymers. Also, the recent development of single molecule experimental techniques has brought along an increase of the interest of physicists in biology.

The mostly studied molecule is definitely DNA, which fascinates by its ability to store the information needed for the synthesis of proteins or RNA. The part of a DNA molecule, which contains the information concerning one protein, is called a gene, while the ensemble of genes in a cell forms the genome. Although genetics is a field in continuous evolution, there are still many open questions regarding DNA: for example, how do cells copy the information stored in DNA during division, or how do they repair DNA? Moreover, recent developments in experimental techniques have made it possible to manipulate DNA in genetic engineering and nanotechnologies, thus creating an interest for the study of DNA properties in conditions that are not necessarily physiologically relevant.

A DNA molecule consists of two polymers of nucleotides that form a double helix. Nucleotides are composed from a phosphate group linked by a phosphoester bond to a sugar ring, which is, in turn, linked to a carbon ring structure called "base". There are four types of bases, which may be part of the DNA structure: adenine (A), guanine (G), thymine (T), and cytosine (C). The first two ones are purines (they contain a pair of fused rings), while the last two ones are pyrimidines (they contain a single ring). The double helical structure of DNA results, on one hand, from stacking interactions between neighboring bases of the same strand and, on the other hand, from hydrogen bonds that form between a purine and a pyrimidine of opposite strands [1]. Adenine and thymine form a double bond, while guanine and cytosine form a triple bond. The breaking of these bonds and the subsequent opening of the double helix is called "denaturation", or "melting", and it can be triggered either thermally, when DNA is heated, or mechanically, when, for example, proteins pull the two strands away from each other.

Besides replication, the best known phenomenon that involves the opening of DNA bases is transcription, that is, the process by which the information contained in a gene is read and used for the synthesis of molecules such as proteins or RNA. Transcription is controlled by so-called

"transcription factors", which are proteins that first connect to the DNA chain at specific sites and then promote transcription by RNA polymerase. In order to initiate this process, RNA polymerase has to recognize and connect to a specific site on double stranded DNA. In eukaryotes, this is done with the help of a few other proteins, the transcription factors, which form a preinitiation complex. The first protein to connect to DNA is the TATA-binding protein, which connects to a specific sequence that is rich in thymine and adenine, called the TATA box, and then mediates the connection of RNA polymerase to the start site of the gene. The TATA-binding protein also opens the DNA double helix by bending it by 80°. Then RNA polymerase catalyses a polymerization reaction, by which it creates an RNA strand one base at a time. At the end of transcription, the newly created RNA molecule is released in the cytoplasm and the RNA polymerase disconnects from the gene.

In fact, most of the processes that take place in a living cell are based on such symbiosis between DNA and proteins. Other notable examples are proteins that are responsible for packing DNA in a cell or for repairing damaged genes. These are all processes that researchers are trying to understand and copy, with the purpose, among others, of developing new generations of medicines for curing genetic diseases.

The first step in the study of all these processes consists in understanding the properties of the DNA double helix and how it interacts with proteins. The study of DNA melting (thermal or chemical) in itself, besides having some interesting practical applications, like genome sequencing and separation, gives useful insight on the mechanisms involved in many of the biophysical processes that include nucleic acids.

The problem when investigating DNA-related phenomena, as well as most other biological processes, is that they are quite difficult to describe in detail by an analytical theory and, because they involve large molecules, they moreover lead to very cumbersome all-atoms simulations. For example, a DNA base pair is composed of about 70 atoms, so that the description of a long DNA chain (hundreds of thousands of base pairs) and the associated buffer molecules involves a very large number of degrees of freedom. This, of course, has a direct consequence on the time interval that can be investigated. For example, the time scale associated with protein folding is of the order of microseconds. Even with today's most powerful computers, simulating such phenomena with all-atoms models would be prohibitively long. The time interval, which is usually simulated with such all-atoms models, is indeed of the order of a few nanoseconds.

Therefore, nowadays' solution for simulating biomolecules, both DNA and proteins, is essentially coarse-grained modeling. This is a technique based on the reduction of the number of degrees of freedom in the system by replacing a group of atoms, like for example a DNA base or a functional group of a protein, by a single particle. The interactions between these particles are modeled by mean-field potentials, which are adjusted to describe the macroscopic properties of the molecule that is specifically studied. The buffer is usually described implicitly by a set of random forces. This permits not only the study of larger systems than by using all-atoms models but also their study at much larger time scales and with longer equilibration periods. One of the first uses of coarse-grained molecular dynamics for the description of proteins goes back to 1975, when Levit and Warshell proposed a model for protein folding [2]. However, it is only recently that coarse-grained modeling has started to be extensively used in the study of biomolecules.

The subject of this work is the study of models of DNA, as well as the interaction between DNA and proteins, at different resolutions. The first part of my thesis concentrates on the investigation of DNA models at different resolutions and how accurately they describe denaturation.

I will start with the simplest models, which are Ising-type (or "statistical") models that describe a base pair by two possible states: open and closed. These models incorporate the effect of both stacking and pairing interactions. The algorithms, which describe melting using such statistical models, have the advantage that they give quite accurate results in a very short computer time. Several free programs are now available, which compute the fraction of DNA open base pairs as a function of temperature (melting curves). These programs are very useful for biological applications, like genome separation, primer design, or PCR. I will first discuss the application of statistical models to the modeling of two-dimensional electrophoresis experiments. Two-dimensional electrophoresis is a method for visualizing polymorphism and comparing genomes, which is based on the separation of DNA fragments subjected to an electric field according first to their size and then to their sequence composition. The first step is based on the fact that the velocities of the DNA fragments in an electric field depend on their size. The second step implies either a temperature gradient or the presence, in increasing concentration, of a chemical denaturant. Using experimental results for the separation of 40 DNA sequences, I will show that a simple expression for the mobility of DNA fragments in both dimensions allows one to reproduce final absolute locations with a precision, which is better than experimental

uncertainties. This part of my thesis was done in collaboration with Bénédicte Lafay, from Laboratoire Ampère (Université de Lyon) and is based on experimental and theoretical work performed in her group [3].

However, there are cases where statistical models are not detailed enough, as for example when it comes to investigate time-dependent properties of DNA. Dynamical models are more efficient for this purpose. By "dynamical model", I mean a model based on explicit expressions for the energy of the DNA sequence written in terms of continuous coordinates and velocities. Such models therefore rely uniquely on the microscopic description of the system and are expected to describe the whole dynamics of DNA, from small vibrations at low temperatures to large amplitude oscillations close to denaturation. The first model, which was designed to study solitons in DNA [4], can however not describe denaturation, because the only degree of freedom for a base pair is its rotation angle around the strand axis. The model that followed (Prohofsky *et al* [5]) considers that the principal source of nonlinearity in DNA are the hydrogen bonds between paired bases (which are usually modeled as Morse potentials), and that the important degree of freedom is the corresponding stretching coordinate. Dauxois, Peyrard and Bishop later replaced, in the model of Prohofsky and coworkers, the harmonic stacking interaction between two successive bases by an anharmonic one, and showed that this leads to denaturation curves that are in better agreement with experiments [6]. Before the beginning of my thesis work, a variant of the Dauxois-Peyrard-Bishop model had been developed in our group, based on the observation that the finiteness of stacking enthalpies is in itself sufficient to insure sharp melting curves [7]. I will present the improvements we brought to this model in order to get still better agreement with experimental results, and show that the improved model provides results that are in quantitative agreement with those obtained from statistical models. Finally, I will describe the critical properties of the model, paying special attention to the narrow region just below the critical temperature.

The second part of this work proposes a model for the description of non-specific protein-DNA interactions and of the search strategies, which DNA-binding proteins use to find their targets. The first studies that prove that proteins are able to bind on a DNA chain were published in 1967 [8,9]. Until then, the general belief was that it is RNA that recognizes sites on DNA rather than proteins. The debate on how proteins connect to DNA started three years later with the experiments of Riggs, Bourgeois and Cohn [10], who measured the *lac* repressor-operator

interaction kinetics and reported an association rate of about 7×10^{9} $M^{-1}s^{-1}$. This is one to two orders of magnitude larger than the rate that was generally assumed for the speed limit of protein-DNA association, if it were to be a purely diffusive process [11,12]. However, this rate was measured in a buffer, the ionic strength of which was much smaller than physiological values. This association rate decreased and became closer to that of diffusion driven reactions once the experiments were repeated at higher salinity. Riggs *et al* therefore concluded that, at low salinities, protein-DNA association must be speeded up by electrostatic attractions between the negative charges on DNA phosphates and positive charges on the protein. Surprisingly enough, most works performed since that time ignore this conclusion of Riggs *et al* and are basically aimed at proposing mechanisms that would enable DNA-protein association to be much faster than normal diffusion.

The target sequence, which must be found by proteins, is very small compared to the size of the whole DNA. For example, a typical target site for the *lac* repressor is composed of about 10-12 base pairs [13], while that of restriction enzymes consists of only 6-8 base pairs [14]. One might therefore wonder how the protein manages to find it in a time scale of about one minute. The generally accepted theory, which is confirmed by recent single molecule experiments, is that a protein connects to any site on DNA through random collisions (non-specific binding), and then searches for its specific site by sliding along the DNA sequence. It then detaches and diffuses again in the buffer if it has not found its target after a certain amount of time. This alternation of sliding and diffusion through the buffer is known as "facilitated diffusion". The specific DNA site differs from non-specific ones by the strength and the nature of its interactions with the protein: non-specific interactions are usually long range and soft (electrostatic), while specific interactions are short-range and sufficiently strong to trap the protein on that particular site (hydrogen bonds)[13].

Most models for the description of protein-DNA interactions that were developed until now are kinetic models of facilitated diffusion [15-19]. They have as ingredients three-dimensional and one-dimensional diffusion coefficients, as well as non-specific association and dissociation rates estimated *a priori*. They also make suppositions about the probability for sliding, and about how many base pairs a protein scans during a single sliding event (sliding length). These assumptions are then used to estimate the expressions of various quantities of interest, such as the association rate and the total time required to find the target, as a function of a set of well-defined

geometric quantities, such as the sequence length L and the cell's volume V. As already mentioned, the more or less implicit goal of most of these models is to show that facilitated diffusion is able to speed up protein-DNA association, whatever the ionic strength of the buffer.

Until now, no coarse-grained dynamical model has been proposed for the description of this phenomenon. In the second part of my thesis I present such a model, which does not involve *a priori* assumptions regarding the motion of the protein. It is based on a description of double stranded DNA, which is inspired from polymer physics [20] and differs from the models studied in the first part of my thesis through the fact that a single spherical bead is used to model fifteen base pairs, thus ignoring the helical shape and the possibility of base pair opening. Moreover, the whole DNA chain is free to move in three dimensions through the cytoplasm. By studying a system formed of a cell containing a protein and several DNA segments, I will show that the proposed model successfully reproduces some of the observed properties of real systems and predictions of kinetic models, like the alternation between three-dimensional diffusion and one-dimensional sliding of the protein along the DNA sequence. Even though these results indicate that this dynamical model indeed displays facilitated diffusion, they also show that its existence does not necessarily imply that the sampling of DNA by proteins happens at rates much larger than the diffusion limit. The most important prediction of this dynamical model is certainly that facilitated diffusion cannot be faster than normal diffusion by a factor larger than two, which is substantially smaller than what is sometimes believed.

I will propose two models for describing the protein: the first one models the protein as a single rigid bead with a charge placed at its center, while the second one assumes that the protein is composed of thirteen beads connected by springs. In the second case, I investigated the association dynamics of both spherical and linear proteins, in order to study the influence of their geometry on the speed of the facilitated diffusion process. I also investigated how other physical properties of the protein, like the charge distribution, elasticity, and position of the search site, affect the DNA sampling process.

At last, I will discuss whether the results obtained with my model are in real contradiction with those of kinetic models, and try to give a clear and well-proofed point of view on protein-DNA association kinetics. I will make a short review of existing experimental and theoretical results and then I will show that, in fact, when using realistic parameters, correct kinetic and dynamical models agree on the issue that facilitated diffusion cannot be much faster than normal

diffusion. For this purpose, I computed the acceleration of targeting due to facilitated diffusion using both types of models and showed that, for experimental values of one-dimensional and three-dimensional diffusion coefficients, such a search strategy is most often even less efficient than normal diffusion.

To conclude this Introduction, I should mention that, at about the same time my first article on this topic was published, some eminent biochemists also expressed the opinion that facilitated diffusion cannot speed up significantly the targeting process. Halford indeed published at the end of 2009 an article entitled “An end to 40 Years of mistakes in DNA-Protein Association Kinetics?”. In this article he contradicts the theory that proteins bind to DNA at rates that surpass the diffusion limit, stating that there is “no known example of a protein binding to a DNA site at a rate above the diffusion limit” [21]. He also points out the fact that Riggs, Bourgeois and Cohn did show that association rates decrease as the ionic strength of the solution increases, and that these results, although validated by subsequent experiments [22-24], have been overlooked for years - and sometimes still are. The results obtained during my PhD work therefore come as a confirmation of these statements, and point out that, indeed, there may have been some longstanding misinterpretations in protein-DNA association kinetics.

Part I: DNA denaturation

2. Short review of DNA melting and DNA models

This chapter contains a short introduction to the structure and properties of DNA (sections 2.1 and 2.2) and a description of the most widely used 1-dimensional statistical (section 2.3) and dynamical (section 2.4) models for DNA melting. It also includes a presentation of the dynamical model of DNA melting that was developed in our group before my arrival, as part of the thesis of Sahin Buyukdagli (section 2.5), as well as a brief sketch of the two methods that were used to investigate its properties, that is Molecular Dynamics simulations and Transfer Integral calculations (section 2.6).

2.1. The structure of DNA

DNA (deoxyribonucleic acid) is the molecule that is responsible for the storage of information of about how, when and where to produce proteins in most living cells. A DNA molecule is a set of two entangled polymers (the "strands"), each strand consisting of a backbone and a chain of bases. The backbone is composed of sugar residues (2-deoxyriboze), which are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These bonds give directionality to the DNA strand, with the ends called 3’ (the one with a terminal hydroxyl group) and 5’ (the one with a terminal phosphate group). A phosphate, a sugar, and the attached base form a nucleotide. There are four types of bases: adenine (A), thymine (T), guanine (G) and cytosine (C). Adenine and guanine are respectively formed by fused five-member and six-member rings (purines), while cytosine and thymine are six-members ring compounds (pyrimidines). The sequence of nucleotides in a DNA strand gives the molecule’s primary structure.

In all living organisms, the two strands are held together by the pairing of complementary bases: as a consequence of both their size and chemical properties, adenine indeed pairs with thymine through hydrogen bonds, while cytosine pairs with guanine. In most cases, all bases of one strand pair with a complementary base on the other strand, so that the genetic information can actually be retrieved from each of the strands. The chemical structure of a DNA double strand is depicted in figure 2.1. Base pairing gives the secondary structure of DNA.

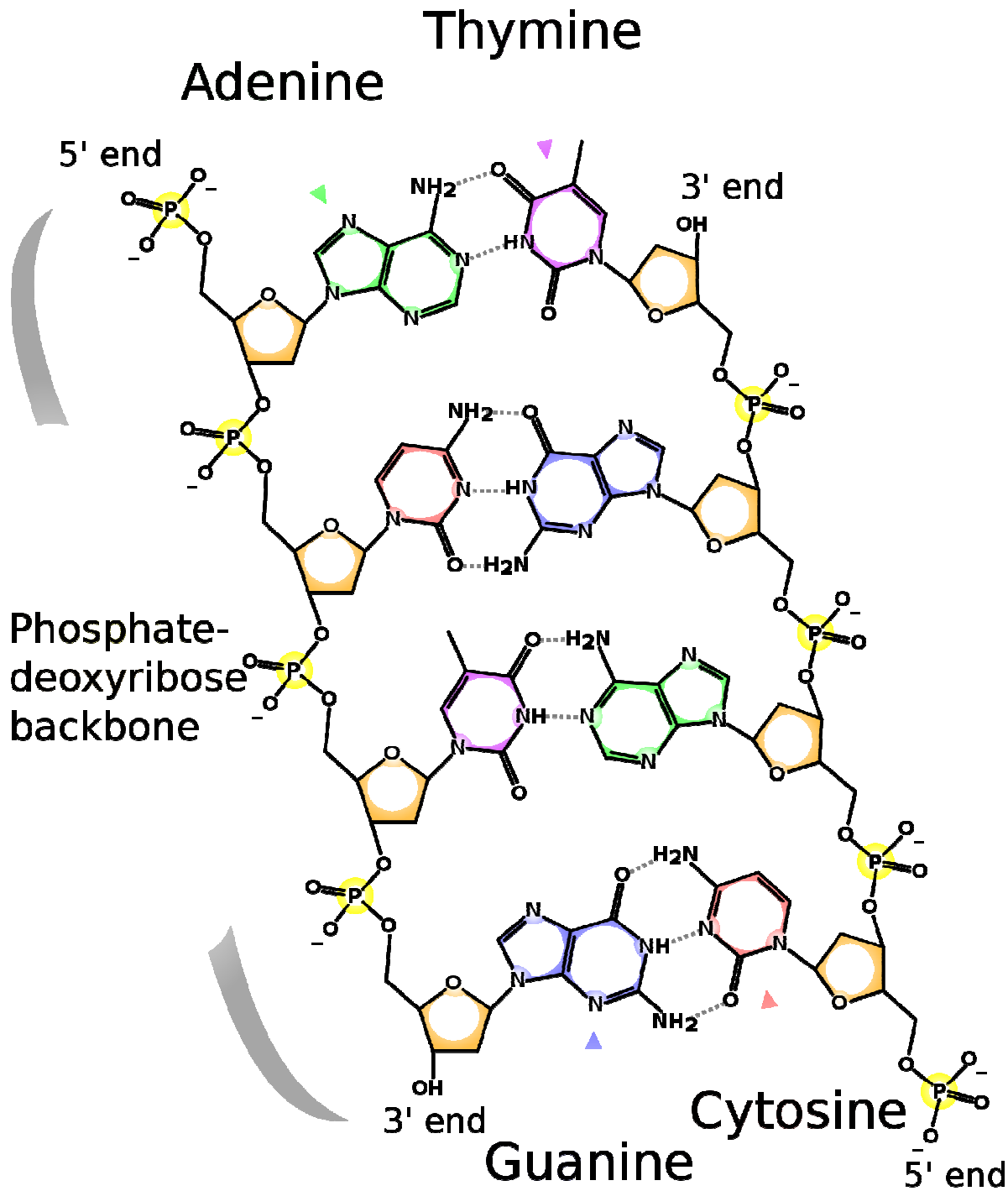

Figure 2.1. *Chemical structure of DNA (secondary structure). Image taken from Wikipedia Commons.*

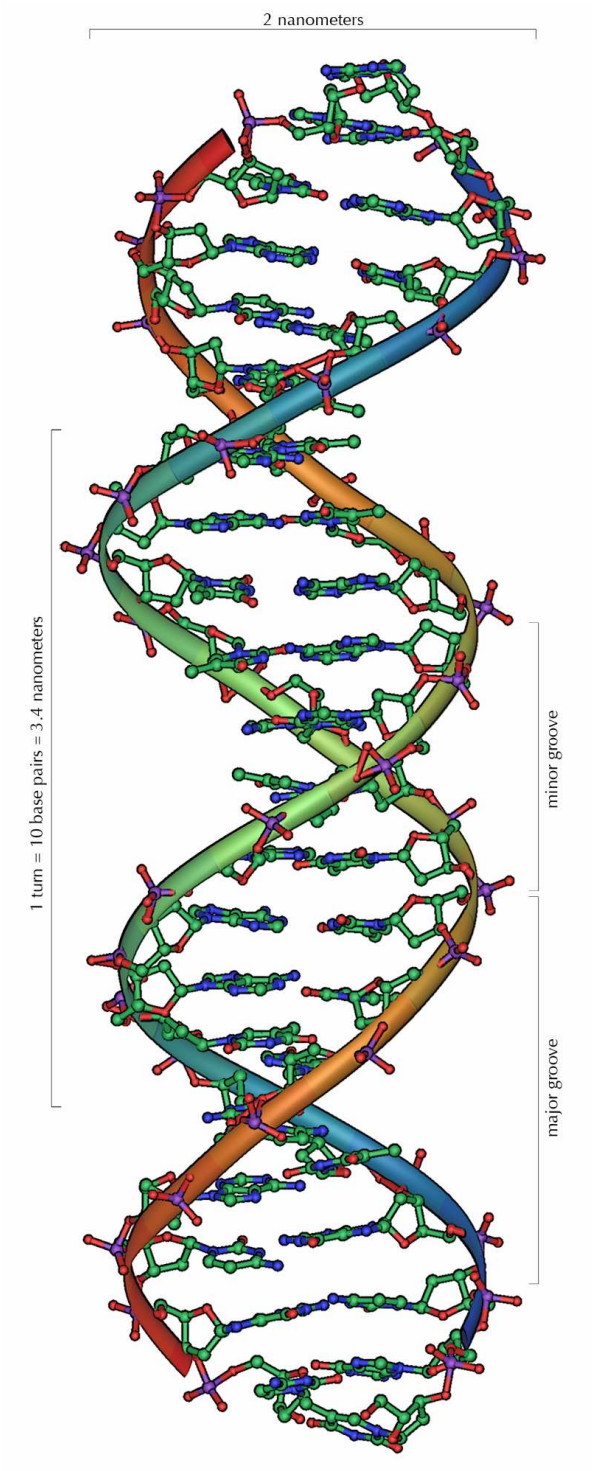

Figure 2.2*. The B-DNA double helix (tertiary structure), with the minor and major grooves highlighted****.*** *Image taken from Wikipedia Commons.*

Moreover, there are forces that act between neighboring bases of the same strand: the stacking interactions. They are due on one hand to attractions between π orbitals of aromatic rings in successive bases and on the other hand to hydrophobic interactions that tend to push bases together. These stacking interactions are responsible for the helical structure of DNA (tertiary structure). There are several possible conformations for the double helix, which differ by the spatial positions of the atoms and the direction of the helix turn, but the one that is ubiquitous in living cells is B-DNA. In this conformation, a turn of the double helix consists of about ten nucleotides. A nucleotide is about 3.3 Å long and the diameter of the double helix is 22 to 26 Å. Actually, the two DNA strands are not perfectly opposite to each other, so the structures form two unequally sized grooves (figure 2.2). The larger one (major groove) is 22 Å wide while the smaller one (minor groove) is 12 Å wide. The major groove is the usual binding site for proteins, because bases are more accessible therein, but there are some proteins that bind into the minor groove (for example, the TATA-binding protein, which has an important role in transcription). The A conformation is shorter and wider than B-DNA, with bases that are tilted rather than perpendicular to the backbone. This structure usually forms in *vitro*, with less water than in physiological conditions, therefore implying weaker hydrophobic interactions. Finally, Z-DNA differs from B-DNA essentially by being a left-handed helix instead of the usual right-handed configuration.

DNA codes the information about protein and RNA structure through the order of the nucleotide sequence along a strand. The parts of DNA that bear information are divided into functional units called genes, which are typically 5000 up to 100000 nucleotides long. Usually, bacteria have about 5000 genes, while humans have about 20000 to 25000, which are eventually separated by long regions, which functions are still not understood. This makes DNA molecules quite long: for example, a human's unpacked genome covers about two meters. A gene usually has two parts: a coding region that specifies the amino acid sequence of a protein and a regulatory region that controls the gene's expression. A single coiled DNA molecule that contains genes, regulatory elements and noncoding regions forms a chromosome. In prokaryotic cells, DNA is found in the cytoplasm while in eukaryotic cells it is located in the nucleus. Actually, in eukaryotic cells, DNA is not free: it is packed around histones, forming a structure called chromatin. The purpose of this packing is to allow the long DNA molecule to fit in a nucleus, to

strengthen DNA to allow for meiosis and mitosis, and to control DNA replication and gene expression.

2.2. DNA melting

DNA melting (also called "denaturation") is the process by witch hydrogen bonds between bases are broken and the two strands separate. These bonds are much weaker than the covalent bonds in the rest of the molecule, so melting does not affect the primary structure of DNA and is a reversible process. Denaturation can be thermal, when DNA is heated, mechanical, if it is caused by a force (for example by proteins that pull one of the strands), or chemical. The most studied one is probably thermal denaturation. Although it differs from the base pairs opening that occurs during transcription, its understanding can still bring a lot of useful information on what happens in this process.

The most widely used method for the experimental study of DNA denaturation is UV absorption spectroscopy: the stacked base pairs in double stranded DNA absorb less ultraviolet light than the bases in a single stranded chain. An increase in temperature causes a sudden opening of base pairs, which is consequently accompanied by an abrupt increase in the absorption, a phenomenon known as hyperchromicity. Therefore, the plots of UV absorption as a function of temperature can be easily compared with theoretically determined plots of the fraction of open base pairs as a function of temperature (melting curves). The temperature at which half of the base pairs in a sequence are open is known as the melting temperature. Its value depends on several factors:

- The types of base pairs the sequence contains: GC pairs are formed of three hydrogen bonds, while AT ones have only two, so the latter will break at a lower temperature.
- The ion concentration of the buffer: positive ions shield the negatively charged phosphates of the backbones. When the ion concentration is small, this shielding is low and the repulsive forces between strands are higher, thus decreasing the melting temperature.
- The presence in the buffer of agents that destabilize hydrogen bonds, such as formamide or urea: these molecules displace hydrates or counterions, having the same phenomenological effect as the decrease in ion concentration.

- The pH: at low pH (acid), the bases become protonated and thus with a positive charge, so they repel each other. At high pH (base), the bases instead lose their protons and they will once again tend to repel each other.

DNA melting is a process that can be assimilated to an order-disorder transition, the order state being given by the paired bases, while the disordered state corresponds to loops formed by broken hydrogen bonds. This analogy is an important point for most of the studies on DNA denaturation. Many of the models that have been developed to describe DNA melting are inspired from the statistical physics of phase transitions (and are therefore named statistical models). They describe a DNA base pair as a spin in the one-dimensional Ising model, that is, as having two possible states: "open", when the hydrogen bond is broken, and "closed", when the hydrogen bond is intact [25,26].

2.3. The Poland-Scheraga statistical model

The types of interactions considered in statistical models are base pairing (free energy G_i) between complementary bases, and stacking (free energy $G^s_{i,i-1}$) between successive bases of the same strand. This leads to a description of the opening of the *i*th base pair as a function of a stability constant s_i, which is expressed as:

$$s_i = \exp\left[-\frac{G_i + G^s_{i,i-1}}{k_B T}\right] \tag{2.1}$$

where T is the buffer's temperature and k_B is Boltzmann's constant. The best known model of this type is certainly that of Poland and Scheraga [27,28]. This model proved to be very efficient in studying the order of the denaturation transition [29] and how it is affected by sequence heterogeneity [30]. In this model the partition function Z_k of a specified state k is a product of three terms: a stability term, which has the form given in equation (2.1), a cooperativity term, which operates between a closed and an open segment, and an entropic term, which takes into account the number of configurations of the denaturated portions of the sequence (loops). This last term was introduced because the nearest-neighbour interactions alone are not sufficient to induce a genuine phase transition (melting just corresponds to a smooth crossover between the closed helix form and the open coil state). This term induces long-range interactions that are

weak but sufficient for a phase transition to occur [27,28]. It is usually taken as a power law of the form:

$$f(m) = (m + D)^{-c}, \tag{2.2}$$

where m is the loop size, D is a stiffness parameter with a generally assumed value of $D = 1$ and the exponent c can take values either of $c = 1.7$ (if it is derived from probabilities of ring closure for self-avoiding random walks) or $c = 2.15$ (if it is estimated using the total number of configurations of a loop embedded in a chain - sharper phase transition). The five mechanisms by which denaturation can propagate and the corresponding terms in the partition function are depicted in figure 2.3.

Poland subsequently provided an algorithm, which allows the efficient calculation of the probability for each base pair to be in the open or closed state [31]. This algorithm works particularly well when combined with the approximation proposed by Fixman and Freiere, which consists in expanding the loop function as an exponential series [32]. Such models were more recently further improved along two directions. First, the various parameters of the model were adjusted against experimental melting curves (for example references [33-35] and references therein). Secondly, it was shown how to take more properly into account excluded volume effects between the loops and the rest of the chain [29,36] (the loop entropy was originally estimated by counting the number of configurations for a closed self-avoiding random walk [37]). This turns out to be of great importance from the physical point of view, since these later calculations lead to a loop closure exponent c greater than 2, which implies that the phase transition is first order,

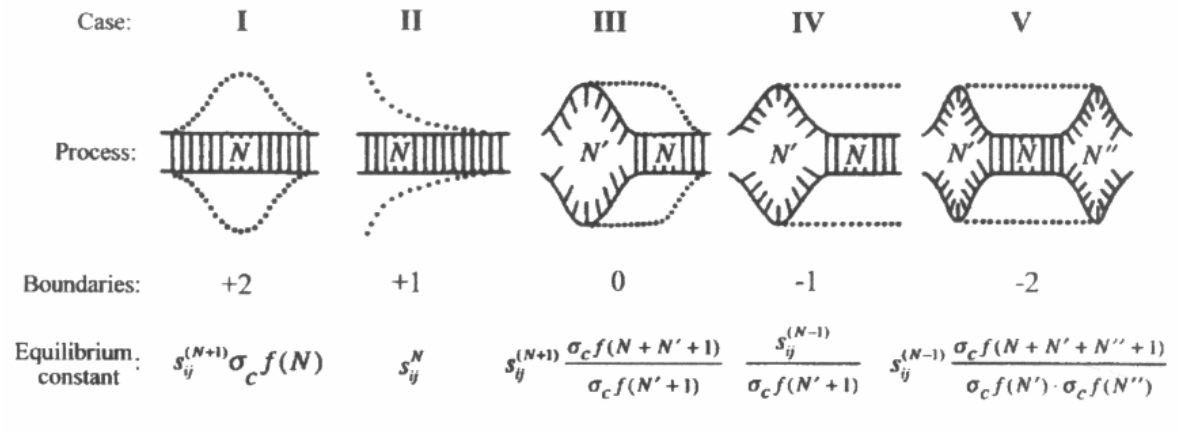

Figure 2.3. *Cartoon of the five mechanisms by which denaturation can propagate in the Poland-Scheraga model, and corresponding terms of the partition function. Image taken from reference [38].*

while the older estimate of c was smaller than 2 and therefore consistent with a second order phase transition.

As a consequence of these improvements, there now exist several online free programs, which provide reliable melting curves of sequences as long as several thousands of base pairs within a few seconds. One of the best known ones is probably Meltsim [38], which is based on Poland's algorithm [31] and Fixman and Freire's speed up approximation [32]. It also gives the user the choice between several parameters sets [33,35,39]. Such a program is of great interest in many areas of biology, like PCR control and mutation analysis.

These programs can also be incorporated in home-written codes for predicting results of various biology experiments, like genome separation through temperature gradient or denaturing gradient electrophoresis display. Electrophoresis is a separation technique, which is based on the fact that, when placed in a gel subjected to an electric field, molecules migrate with different speeds according to their size and charges. For writing the corresponding code, one does not need to fully understand the algorithms behind the computation of denaturation curves, but only to have a tool that is fast, reliable and easy to use. In turn, this kind of simulations help optimize experimental conditions (denaturing gradient or temperature range, electrophoresis duration) without having to perform a large number of tedious preliminary experiments, and to predict whether electrophoresis is a convenient tool to identify a given mutation or a difference that is expected to exist between the genomes of closely related organisms. Chapter 3 of this report precisely contains a detailed study of the modelling of two-dimensional DNA electrophoresis separation experiments.

However, statistical models are not always detailed enough, especially when it comes to describing time dependent phenomena. In this case, a more suitable approach is that provided by dynamical models. I briefly describe dynamical models in the two following sections (sections 2.4 and 2.5). I moreover describe in chapter 4 some work aimed at optimizing the parameters of the dynamical model developed in our group, and I compare the melting curves computed therewith with experimental ones and those obtained from statistical models.

2.4. The Dauxois-Peyrard-Bishop dynamical model

Dynamical models are based on explicit expressions for the energy of the DNA sequence, which are written in terms of usual continuous coordinates and velocities. They are conceptually appealing in the sense that one just needs to provide a microscopic description of the system, like kinetic energy and the shape of pairing and stacking interactions. Its macroscopic properties and evolution with temperature then unequivocally follow there from. Stated in other words, if the masses and characteristic energies introduced in the Hamiltonian are reasonable and the derived macroscopic properties match experimental results, then one might feel confident that the microscopic description of the system is correct. Moreover, dynamical models are of course mandatory as soon as one is not interested in averaged quantities but rather in transient phenomena and fluctuations [40].

The first dynamical models for the description of DNA were developed for the study of soliton wave propagation in the DNA double strand. In 1980, hydrogen-deuterium exchange experiments evidenced the propagation of base pairs openings along the chain in a manner that resembles that of solitons [41]. In the same work the authors proposed a Hamiltonian for describing DNA:

$$H = \sum_n \left\{ \frac{1}{2} m r^2 \dot{\phi}_n^2 + \frac{K}{2} (\phi_n - \phi_{n-1})^2 + mgr(1 - \cos\phi_n) \right\} \tag{2.3}$$

where r is the length of the bond (considered rigid) between the base and the sugar-phosphate backbone and ϕ_n is the rotation angle of the basis around the strand axis. The first term in the Hamiltonian denotes the kinetic energy, the second one gives the torsional elastic energy of the bases, while the third one is an attractive potential that describes hydrogen bonds between bases. This first model is quite simple, but further on more complex models have been developed [42-47]. However, these models are not suitable to describe DNA thermal denaturation because the only degree of freedom is the rotation of the bases around the strands, while denaturation is better described by the stretching of the base pairs. To my knowledge, Prohofsky and co-workers were the first to propose a dynamical model that describes DNA denaturation [5,48]. The Hamiltonian of this model writes:

$$H = \sum_n \left\{ \frac{m}{2}\left(\dot{u}_n^2 + \dot{v}_n^2\right) + \frac{K}{2}\left[\left(u_n - u_{n-1}\right)^2 + \left(v_n - v_{n-1}\right)^2\right] + V_M\left(u_n - v_n\right) \right\} \tag{2.4}$$

where u_n and v_n are the displacements from the equilibrium positions of the two bases that compose the nth base pairs, taken along the axis that is perpendicular to the backbone and joins the two strands. The first term gives the kinetic energy, the second term contains the stacking interaction between successive bases, which is considered here to be harmonic, and, finally, the last term describes the base-pairing interactions as Morse potentials:

$$V_M(u_n - v_n) = D(1 - e^{-a(u_n - v_n)/\sqrt{2}})^2 \tag{2.5}$$

By making a change of variables from the absolute coordinates u_n and v_n to symmetric and antisymmetric coordinates:

$$\begin{aligned} x_n &= (u_n + v_n)/\sqrt{2} \\ y_n &= (u_n - v_n)/\sqrt{2} \end{aligned} \tag{2.6}$$

the Hamiltonian can be rewritten as:

$$H = H_1(x_1, x_2, ..., x_n) + H_2(y_1, y_2, ..., y_n) \tag{2.7}$$

where the first term describes the motion of the centers of mass of the base pairs:

$$H_1 = \sum_n \left\{ \frac{m}{2}\dot{x}_n^2 + \frac{K}{2}(x_n - x_{n-1})^2 \right\} \tag{2.8}$$

while the second terms describes the forming and breaking of base pairs:

$$H_2 = \sum_n \left\{ \frac{m}{2}\dot{y}_n^2 + \frac{K}{2}(y_n - y_{n-1})^2 + D(1 - e^{-ay_n})^2 \right\} \tag{2.9}$$

This Hamiltonian also has solitonic wave solutions, which have been studied in detail [49,50].

However, this model does not accurately describe the denaturation transition. Dauxois, Peyrard and Bishop (DPB) indeed pointed out that it leads to a much too smooth denaturation process, but that this is no longer the case upon introduction of an anharmonic stacking interaction [6]:

$$H_2 = \sum_n \left\{ \frac{m}{2}\dot{y}_n^2 + \frac{K}{2}\left(y_n - y_{n-1}\right)^2\left[1 + \rho e^{-\alpha(y_n + y_{n-1})}\right] + D\left(1 - e^{-ay_n}\right)^2 \right\} \tag{2.10}$$

This new expression implies that the stacking interaction becomes weaker when the corresponding base pairs separate further, thus decreasing the stiffness of the chains and leading to a much steeper phase transition [6]. It is based on the hypothesis that, when the hydrogen

bonds connecting the bases break, the electronic distribution on the bases is sufficiently modified to let the stacking interaction between the bases decrease significantly.

The model of equation (2.10), which from now on will be referred to as the DPB model, was essentially used to investigate the dynamics of short sequences containing from several tens to a few hundreds base pairs [51,52], but it was shown in our group that it is also able to reproduce the characteristic peaks, which appear in the melting curves of inhomogeneous sequences in the 1000-10000 base pairs range [53] (the peaks that were reported for periodic DNA sequences with two or three base pairs in the unit cell [54] essentially arise from end effects and are not directly related to the experimentally observed ones).

2.5. The dynamical model developed in our group

A few years ago, a variant of the DPB model was developed in our group [7], which is closer to statistical ones than the DPB model, in the sense that it is based on site-specific, finite stacking enthalpies (numerical values for the enthalpies were borrowed from table 1 of [38]). It is based on the observation that the finiteness of the stacking interaction is in itself sufficient to insure a sharp melting transition. Since I will report in chapter 4 on several results obtained with this model, I now describe it in some detail.

The general form of the Hamiltonian is:

$$
\begin{aligned}
&H = E_{\mathrm{kin}} + V \\
&E_{\mathrm{kin}} = \frac{m}{2}\sum_{n=1}^{N}\left(\frac{\mathrm{d}y_n}{\mathrm{d}t}\right)^2 \\
&V = \sum_{n=1}^{N} V_M^{(n)}(y_n) + \sum_{n=2}^{N} W^{(n)}(y_n, y_{n-1}) \\
&V_M^{(n)}(y_n) = D_n\left(1-\exp(-a\,y_n)\right)^2 \\
&W^{(n)}(y_n, y_{n-1}) = \frac{\Delta H_n}{C}\left(1-\exp\left(-b(y_n - y_{n-1})^2\right)\right) + K_b(y_n - y_{n-1})^2
\end{aligned}
\qquad (2.11)
$$

where N is the number of base pairs in the sequence and y_n a measure of the distance between the paired bases at position n. More precisely, if u_n and v_n denote the displacements of the two bases of pair n from their equilibrium positions along the direction of the hydrogen bonds that connect them, then y_n is defined as in equation (2.6), that is $y_n = (u_n - v_n)/\sqrt{2}$ [5,55]. In the

expression for the kinetic energy E_{kin}, m denotes the mass of a nucleotide, which we assume to be independent of the precise nature of the base pair at position n (numerically, we use m=300 amu). As for the DPB model, the potential energy V is the sum of two different contributions, namely on-site potentials $V_M^{(n)}(y_n)$ and nearest-neighbour interaction potentials $W^{(n)}(y_n, y_{n-1})$. $V_M^{(n)}(y_n)$ represents the two or three hydrogen bonds that connect the paired bases at position n and is taken as a Morse potential of depth D_n, as in the original model of Prohofsky and coworkers [5,48]. $V_M^{(n)}(y_n)$ is often called a "pairing" potential, because it is an increasing function of $|y_n|$ and therefore opposes the dissociation of the pair. Our model differs from the DPB one essentially in the $W^{(n)}(y_n, y_{n-1})$ interaction, which is again the sum of two terms, namely the stacking potential plus the backbone stiffness. Both terms are increasing functions of $|y_n - y_{n-1}|$, which means that they oppose the de-stacking of the bases, that is, the separation of successive bases belonging to the same strand. The stacking potential is modelled by a gaussian hole of depth $\Delta H_n / C$, while the backbone stiffness is taken as a harmonic potential of constant K_b. Its role consists in preventing dislocation of the strands, that is, in insuring that bases belonging to the same strand do not separate infinitely when approaching the melting temperature.

The numerical values for the parameters of equation (2.11) used in previous works [7,40,53,56-58] were obtained in the following way: the ten stacking enthalpies ΔH_n were borrowed from statistical models (table 1 of reference [38]) and it was assumed that the paired bases do not unstack simultaneously, which implies that $C = 2$ [7]. On the other hand, a uniform stacking strength of $\Delta H_n / C = 0.22$ eV was used to model the homogeneous sequences that are involved in most statistical studies. D_n =0.04 eV and a=4.45 Å^{-1} were taken from the DPB model [6], while K_b =10^{-5} eV Å^{-2} was fixed somewhat arbitrarily. Finally, b was varied to get a 40 to 50 K separation between the melting temperatures of pure AT and GC sequences, as in experiments performed at physiological salinities and pH values. It was consequently fixed at b=0.10 Å^{-2}.

Actually, the results obtained with this set of parameters should be improved with respect to at least three points. First, the denaturation curves have a weak temperature resolution, in the sense that they are somewhat too smooth compared to the curves obtained with statistical models or experimental ones. Moreover, the critical temperature diminishes much too quickly with decreasing sequence lengths. As reported in [57], the lowering of the melting temperature

behaves as 3250/*N* for this model and 1850/*N* for the DPB model, while most online oligonucleotide property calculators assume a 500/*N* dependence (which agrees with the experimental results reported in [59]) and statistical models even predict a gap smaller than 1 K between the melting temperatures of an infinitely long homogeneous sequence and its finite counterpart with *N*=100 base pairs. Finally, mechanical unzipping experiments performed at constant force show that the critical force, which is needed to keep the two strands of a DNA sequence open at around 20°C, lies in the range 10-20 pN [60,61], while this model predicts that a few pN are sufficient. As will be argued below, this poor agreement between predicted and measured critical forces is essentially ascribable to a too small value of K_b (remember that this parameter was fixed arbitrarily), while the exaggerated sensitivity of the melting temperature with sequence length results from the too large depth of the stacking interaction. I will show in chapter 4 that it is sufficient, once these two points have been corrected, to slightly adjust the remaining parameters in order to reproduce experimental denaturation curves more correctly.

2.6. Molecular Dynamics simulations and Transfer Integral calculations

The two methods, which were used in our group to investigate the properties of dynamical models, are Molecular Dynamics (MD) simulations and Transfer-Integral (TI) calculations. MD simulations consist in integrating step by step the Langevin equations:

$$m\frac{\mathrm{d}^2 y_n}{\mathrm{d}t^2} = -\frac{\partial H}{\partial y_n} - m\gamma\frac{\mathrm{d}\, y_n}{\mathrm{d}t} + w(t)\sqrt{2m\gamma k_B T} \qquad (2.12)$$

with a second order Brünger-Brooks-Karplus integrator [62]. γ is the dissipation coefficient (we assumed $\gamma = 5$ ns^{-1}) and $w(t)$ a normally distributed random function with zero mean and unit variance. The second and third term in the right-hand side of equation (2.12) model the effect of the buffer on the DNA sequence. The sequence is first heated by subjecting it to a temperature ramp, which is slow enough for the physical temperature of the system (calculated from the average kinetic energy) to follow the temperature of the random kicks (the symbol *T* in equation (2.12)). The average values of the quantities we are interested in are then obtained by integrating Langevin equations at constant temperature for time intervals of 100 ns.

MD simulations are very easy to implement, but they have two limitations: firstly, they require a very large amount of CPU time, because step by step integration of hundreds or

thousands of coupled differential equations is intrinsically slow, and secondly, the temperature resolution of the results is rather poor, especially close to the melting temperature, because of the very slow fluctuations of temperature of the sequence in this range [40].

On the other side, the TI method [54,63] is a mathematical technique to replace the *N*-dimensional integrals, which appear for example in the expressions of the partition function *Z*

$$Z = \int \mathrm{d}y_1 \, \mathrm{d}y_2 \ldots \mathrm{d}y_N \exp(-\beta V). \tag{2.13}$$

and the average bond length at position *n*

$$\langle y_n \rangle = \frac{1}{Z} \int \mathrm{d}y_1 \, \mathrm{d}y_2 \ldots \mathrm{d}y_N \; y_n \exp(-\beta V) \tag{2.14}$$

by products of *N* one-dimensional integrals. Other quantities of interest, like the free energy per base pair, *f*, the entropy per base pair, *s*, and the specific heat per base pair, c_V, are then easily obtained from *Z* according to

$$\begin{aligned} f &= -\frac{1}{N\beta} \ln(Z) \\ s &= -\frac{\partial f}{\partial T} \\ c_V &= -T \frac{\partial^2 f}{\partial T^2} \end{aligned} \tag{2.15}$$

Note that we use finite differences for the calculation of *s* and c_V. When it works, the TI method is very efficient, in the sense that it enables to calculate most quantities much more rapidly and with a better temperature resolution than MD simulations. As discussed in some detail by Zhang *et al* [54], the TI kernel is however singular when using a bound on-site Morse potential, so that one needs to check carefully the convergence of the obtained results with respect to the upper bound for *y*, which is assumed in practical calculations. A general observation would be that, at the thermodynamic limit of infinitely long homogeneous chains, there always exists a certain temperature range surrounding the critical temperature, where the TI method is not valid. For some sets of parameters, this interval is so large that the TI method is essentially useless (this is, of course, not the case for the set of parameters that will be proposed in chapter 4). In contrast, calculations are more reliable for finite sequences, because they melt at temperatures that are lower than the critical temperature of the infinite sequence.

I have not been involved in the development of the TI codes. Therefore, I do not provide here more detail on this technique. More precisions can be found, for instance, in the work of Zhang *et al* [54] and in some publications of our group [56-58].

In order to illustrate the capabilities of the two methods, the temperature evolution of the average base pair separation, $\langle y \rangle = \frac{1}{N} \sum \langle y_n \rangle$, and of the average pairing and potential energy per base pair, ($u = \langle V \rangle / N$), are shown in figures 2.4 and 2.5, respectively. Solid lines show results obtained from TI calculations, and dashed lines results obtained from MD simulations, for a homogeneous sequence with 1000 base pairs. It can be seen that the agreement between both types of calculations is generally very good, except close to the critical temperature, where MD simulations are much noisier than TI calculations (although 10 trajectories were averaged, so that MD simulations required between 10 and 100 times more CPU time than TI calculations) and evolve less sharply with temperature. As mentioned above, this difference is due in part to the very slow temperature fluctuations of the sequence in this interval [40], and in part to the fact that the averaging time between two temperature increments (100 ns per K) is too small compared to the characteristic times of the denaturation dynamics of the sequence.

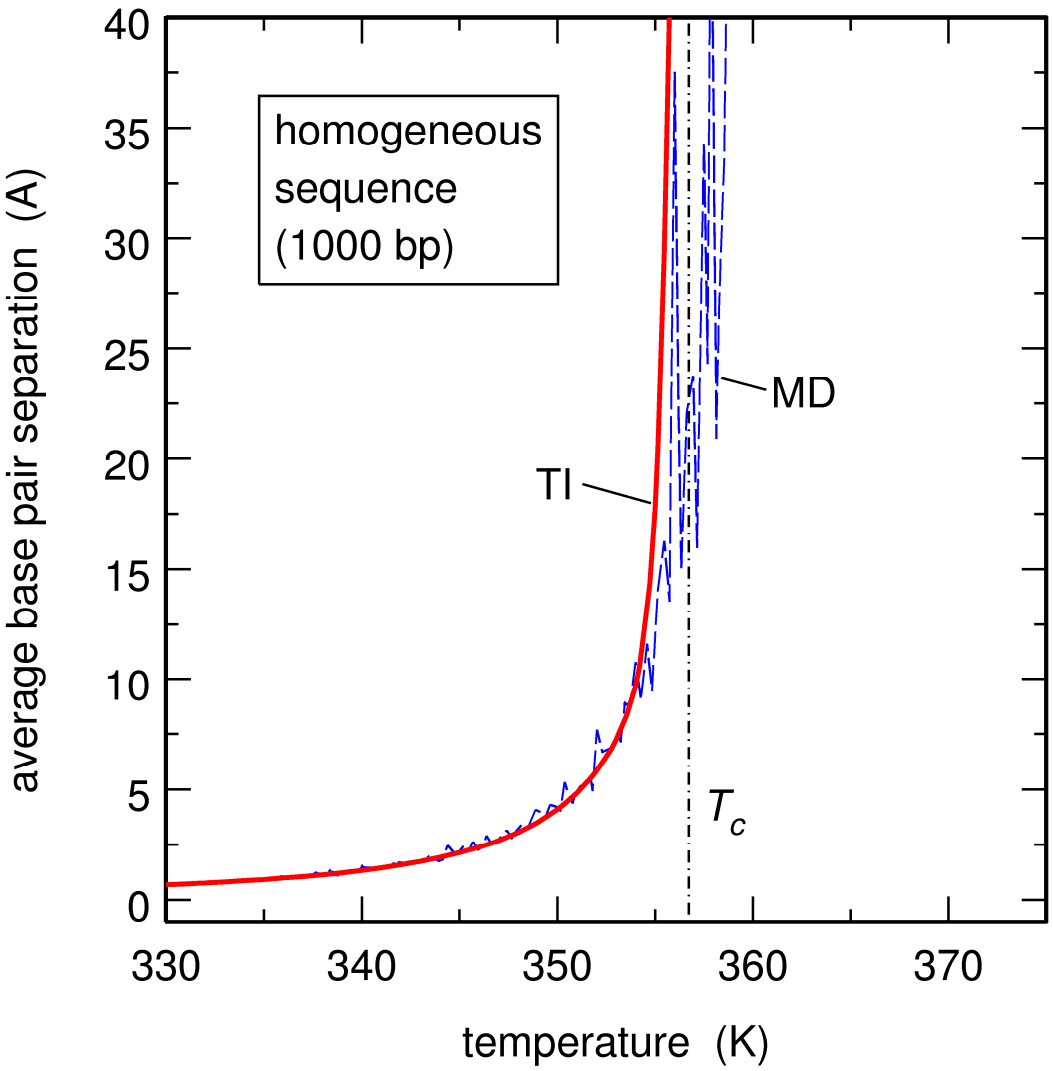

Figure 2.4. *Plot, as a function of the temperature T of the sequence, of the average base pair separation* $\langle y \rangle = \frac{1}{N}\sum_{n=1}^{N}\langle y_n \rangle$ *for a homogeneous sequence with 1000 base pairs, obtained from MD simulations (dashed line) and TI calculations (solid line). Calculations were performed for the model in equation (2.11) and the parameters reported in section 4.3.* $\langle y \rangle$ *is expressed in Å. The vertical dot-dashed line shows the critical temperature for this sequence (*$T_c(N) = 356.73$ *K). Each point of the MD curve corresponds to a total accumulation time of 1* μ*s.*

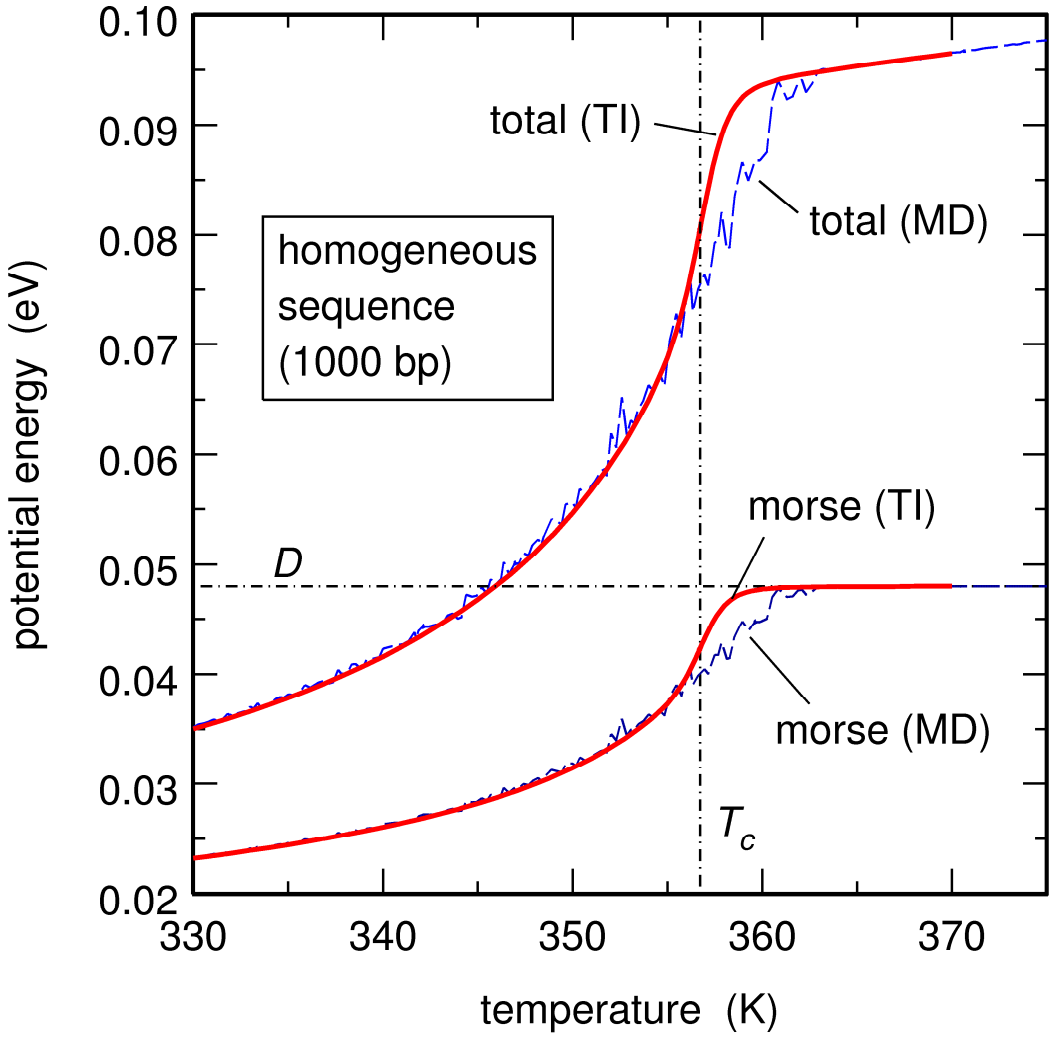

Figure 2.5. *Plot, as a function of the temperature T of the sequence, of the average energy in each Morse oscillator and the average total potential energy per base pair,* $u = \langle V \rangle / N$ *, for a homogeneous sequence with 1000 base pairs, obtained from MD simulations (dashed lines) and TI calculations (solid lines). Calculations were performed for the model in equation (2.11) and the parameters reported in section 4.3. Energies are expressed in eV. The vertical dot-dashed line shows the critical temperature for this sequence (*$T_c(N) = 356.73$ *K). Each point of the MD curve corresponds to a total accumulation time of 1 μs. For TI calculations, u was obtained from equation (2.15) and the relation* $u = f + T s$ *.*

3. Application of statistical models to 2D electrophoresis display

In this chapter, I will describe how I built a program around MeltSim [38] (one of the free programs for computing DNA melting curves that are based on the Poland algorithm) to get predictions of the results of electrophoresis experiments, and I will show that these experiments can indeed be modelled with high accuracy. As already mentioned in the Introduction, this part of my thesis has been done at the suggestion of Bénédicte Lafay from Laboratoire Ampère (Université de Lyon), and it is based on experimental and theoretical work [3] performed in her group.

3.1. Introduction

Electrophoresis is a separation technique that is widely used by biologists. It is a fast and economical way of visualizing polymorphism and comparing genomes. An interesting variation of this rather old technique is two-dimensional (2D) DNA display, which was first described by Fisher and Lerman [64-66]. It consists in separating DNA fragments in two steps, first according to their size and then to their sequence composition. The first step uses traditional slab electrophoresis, for example in agarose or polyacrylamide gels. Collisions between DNA and the gel reduce the mobility of DNA fragments, so that the gel acts as a sieve and the electrophoretic mobility becomes size-dependent, with smaller molecules generally going faster than large ones [67]. In the second dimension, fragments of identical lengths are separated on the basis of their sequence composition, thanks to a gradient of either temperature (TGGE : temperature gradient gel electrophoresis) or concentration of a chemical denaturant present in the buffer, e.g., a mixture of urea and formamide (DGGE : denaturing gradient gel electrophoresis), both methods being closely related [68,69]. The effective volume of denaturated regions of DNA being larger than that of double-stranded ones, the mobility of a given fragment decreases as the number of open base pairs increases. Since AT-rich regions melt at lower temperatures than GC-rich ones, GC-rich fragments usually move farther than AT-rich ones.

Although 2D DNA display has already been applied to the comparison of the genomes of closely related bacteria [70-72], this method is still essentially empirical and simulations have only been used to a very limited extent to plan experiments and analyze results [3,73,74]. In particular, it has been shown only recently [3] that separation of DNA fragments in 2D display

experiments can be predicted with satisfying precision using a model that combines step-by-step integration of the equations of motion of each fragment and the use of the open source program MeltSim [38] to estimate the number of open base pairs at each step of the DGGE phase. In this work the method was validated by predicting the outcome of the separation of 40 sequences. The first separation was done in a 0.8% agarose gel in 40mM Tris, 1mM Na_2EDTA at 2 V/cm for 8h. For the second dimension, a 4% polyacrylamide gel with parallel ascending gradient of formamide (10-40%) and urea (1.8-7M) was used and the separation lasted for 24h, at a constant temperature of 60°C and in an electric field of 7 V/cm. The setups used for two separations are shown in figures 3.1 and 3.2.

However, in reference [3] were not used the computed absolute final positions of the DNA sequences, but only their positions relative to two reference segments, because the absolute positions were wrong by more than 1 cm (that is, several tens of percents of the total displacement). The fact that the errors in computed absolute final positions are so large is worrying in itself, because it unambiguously indicates that something is wrong in the model. Moreover, using relative positions is quite dangerous, since an eventual error in the coordinates of one of the reference sequences implies that the positions of all other sequences will be wrongly predicted. Also, some of the expressions that were used to compute the mobility of the fragments are quite imprecise and depend on too many parameters.

The goal of the work presented in this chapter was to extend the results presented in reference [3] along several lines. I have used absolute coordinates, instead of relative ones, and shown that they can be computed with uncertainties that are smaller than experimental ones. I have also shown that one can use simple expressions to describe correctly the mobility of a DNA sequence in a gel.

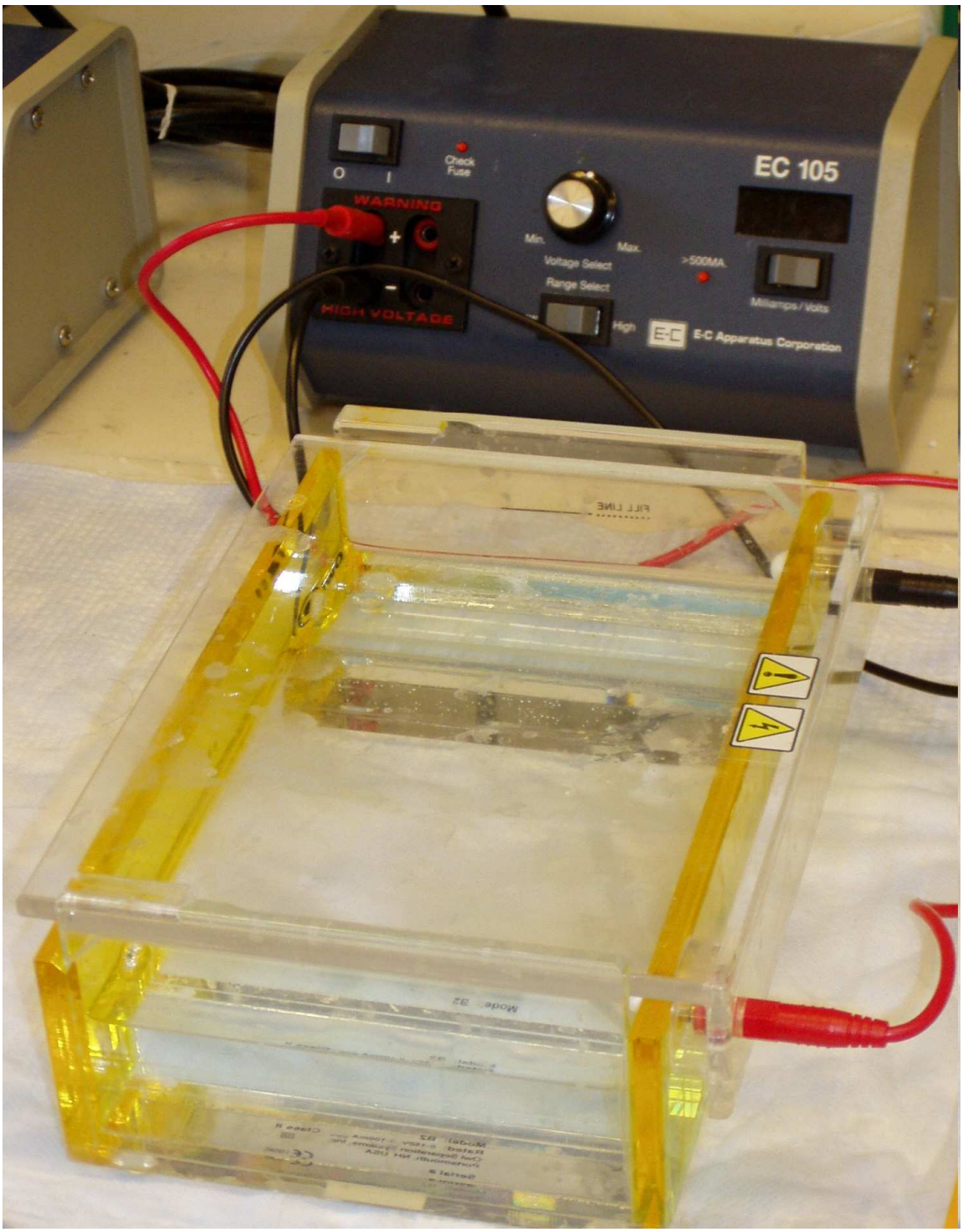

Figure 3.1. The experimental setup used for the separation in the first dimension (horizontal agarose gels).

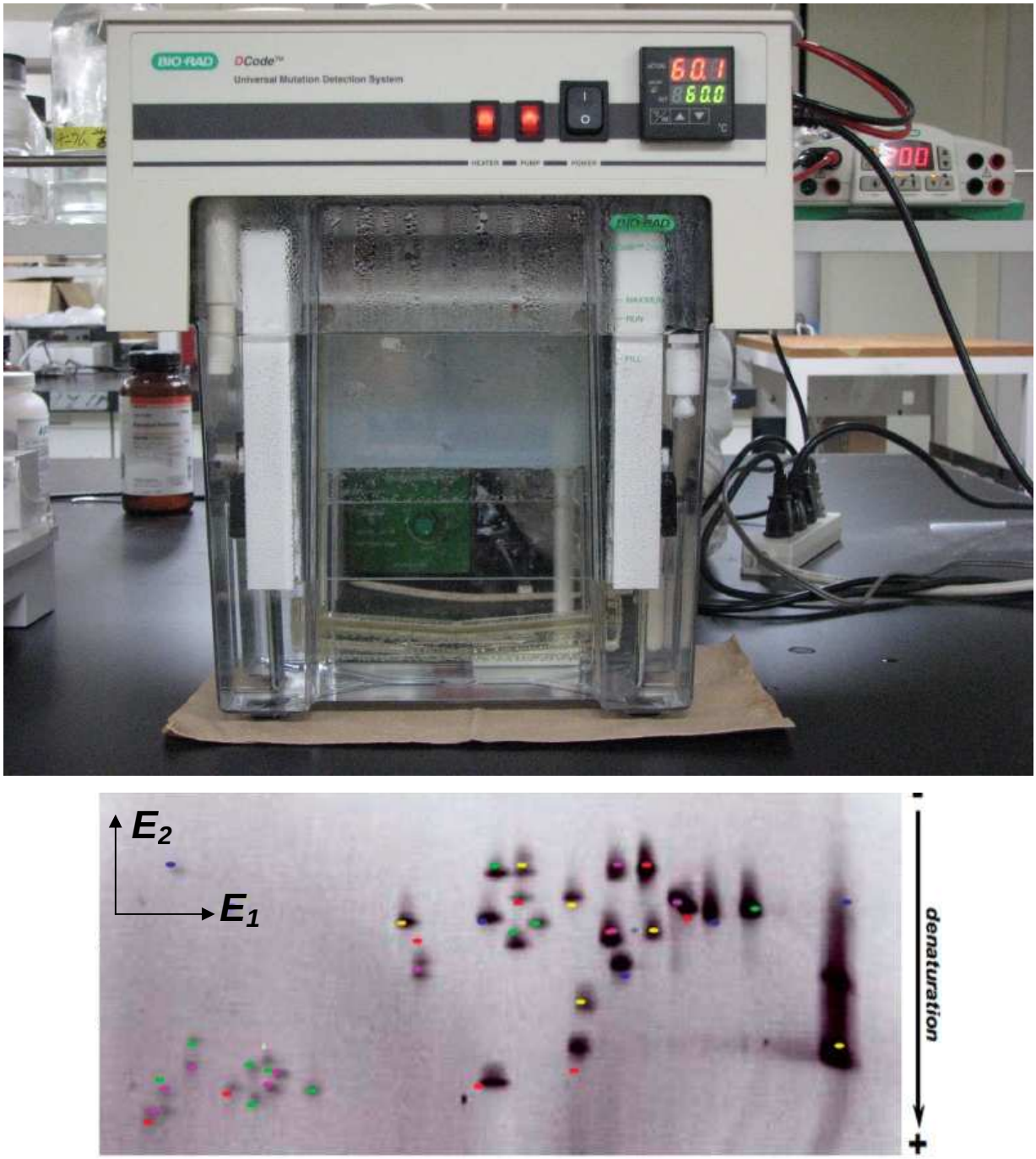

Figure 3.2 *Upper image: The setup used for the separation experiments in a second dimension (Dcode Universal Detection System-vertical polyacrylamide gels) Lower image: Experimental display of the sequences discussed here (a gel shown at the end of separation in the two dimension). Super-imposed in colors are the simulation results of reference [3]. The color code is the following: yellow for EcoRI digested sequences, red for Eco91, pink for Eco47I, blue for HindII and green for PstI.*

3.2. General framework

According to the definition of mobility, the position *y* of sequence *s* at time *t* in a constant electric field *E* satisfies the relation:

$$\frac{dy}{dt} = \mu(s, y)E \tag{3.1}$$

If the mobility $\mu(s, y)$ depends uniquely on the sequence *s* and not on position *y*, as is the case for the standard electrophoresis set-up in the first dimension, integration of equation (3.1) is straightforward and leads to

$$y(t) - y(0) = \mu(s)E\,t \tag{3.2}$$

In contrast, if the mobility $\mu(s, y)$ depends on both the sequence *s* and position *y*, as is the case for TGGE and DGGE, then equation (3.1) must be integrated step by step, according to

$$y(t + dt) = y(t) + \mu(s, y(t))E\,dt \tag{3.3}$$

Here, I integrated such equations of motion for the same 40 DNA fragments discussed in reference [3] using the same conditions as described therein. These fragments were obtained from the site-specific restrictions of λ-phage genomic DNA using EcoRI, Eco47I, Eco91I, HindIII and PstI, respectively (however, here it is not important how the DNA sequence were obtained, but it is only their sizes and base compositions that matter). As reported in the first columns of table 3.1, the size of these fragments varies between 1929 and 23130 base pairs, and their GC content between 36.0% and 58.9%. For the first separation (according to size), I plugged the experimental values of the electric field (E=2 V/cm) and the total electrophoresis time (t=8 h) in equation (3.2). For the second separation (according to sequence composition), equation (3.3) was integrated with the experimental value E=7 V/cm for 44 h by steps of 7 minutes. I also checked that results do not vary when the total integration time is increased to 80 h and the time step lowered to 1 minute. These results are similar to those obtained with an integration time of 24 h, which coincides with the experimental duration, showing that DNA sequences were already stopped at the end of the electrophoresis experiments.

			1st dimension				2nd dimension			
Enzyme	Length (bp)	GC %	y_{exp} (cm)	σ_{exp} (cm)	y_{calc} (cm)	Δy (cm)	y_{exp} (cm)	σ_{exp} (cm)	y_{calc} (cm)	Δy (cm)
EcoRI	21226	56.9	2.45	0.04	2.35	0.10	3.56	0.21	3.57	-0.01
	7421	44.5	5.02	0.10	5.09	-0.07	2.39	0.27	2.41	-0.02
	5804	49.6	6.03	0.14	6.08	-0.05	3.10	0.19	3.11	0.00
	5643	43.2	6.20	0.18	6.20	0.00	2.07	0.31	2.18	-0.11
	4878	39.7	6.89	0.18	6.87	0.02	1.84	0.35	1.76	0.08
	3530	44.0	8.61	0.21	8.54	0.07	2.36	0.31	2.47	-0.10
Eco47I	8126	47.8	4.68	0.10	4.76	-0.08	2.07	0.32	2.29	-0.22
	6555	38.0	5.56	0.12	5.57	-0.01	1.81	0.34	1.75	0.07
	6442	43.7	5.61	0.12	5.64	-0.03	2.45	0.29	2.42	0.03
	3676	47.5	8.35	0.20	8.32	0.03	2.70	0.25	2.96	-0.25
	2606	56.4	10.31	0.21	10.29	0.02	3.89	0.18	3.85	0.04
	2555	56.7	10.45	0.25	10.41	0.04	4.04	0.17	4.00	0.04
	2134	55.3	11.58	0.27	11.52	0.06	3.83	0.21	3.83	0.01
	2005	57.6	11.87	0.27	11.92	-0.05	4.30	0.18	4.00	0.30
	1951	58.5	12.03	0.28	12.09	-0.06	4.38	0.17	4.25	0.13
Eco91I	8453	46.7	4.52	0.08	4.62	-0.10	2.17	0.30	2.43	-0.25
	7242	47.1	5.13	0.12	5.18	-0.05	1.79	0.34	1.76	0.04
	6369	46.0	5.68	0.14	5.69	-0.01	2.44	0.26	2.48	-0.04
	5687	56.4	6.12	0.16	6.17	-0.05	3.51	0.16	3.77	-0.26
	4822	40.2	6.96	0.19	6.93	0.03	2.08	0.32	2.12	-0.04
	4324	58.1	7.46	0.18	7.46	0.0	3.88	0.13	3.94	-0.06
	3675	46.0	8.42	0.20	8.32	0.10	2.84	0.25	2.72	0.12
	2323	57.8	11.06	0.24	10.99	0.07	4.06	0.18	4.01	0.05
	1929	58.9	12.09	0.28	12.16	-0.07	4.33	0.19	4.29	0.04
HindIII	23130	55.9	2.32	0.04	2.22	0.10	2.81	0.30	2.25	0.56
	9416	45.0	4.18	0.09	4.27	-0.09	2.20	0.31	2.37	-0.17
	6682	48.0	5.49	0.12	5.49	0.00	2.70	0.24	2.83	-0.13
	4361	45.2	7.37	0.17	7.42	-0.05	2.30	0.30	2.46	-0.15
	2322	37.1	10.97	0.26	11.00	-0.03	2.36	0.37	2.29	0.07
	2027	36.0	11.82	0.26	11.85	-0.03	2.00	0.41	1.76	0.24
PstI	11497	46.8	3.63	0.06	3.68	-0.05	2.20	0.32	2.29	-0.09
	5077	44.9	6.69	0.16	6.68	0.01	2.32	0.29	2.50	-0.18
	4749	43.8	7.02	0.19	7.00	0.02	2.51	0.26	2.48	0.04
	4507	36.0	7.31	0.18	7.26	0.05	1.86	0.34	1.77	0.09
	2838	56.6	9.85	0.24	9.78	0.07	4.02	0.16	4.07	-0.05
	2560	53.2	10.39	0.22	10.40	-0.01	3.74	0.18	3.76	-0.02
	2459	57.7	10.61	0.24	10.64	-0.03	4.15	0.17	4.11	0.04
	2443	54.8	10.69	0.24	10.68	0.01	3.84	0.19	3.82	0.02
	2140	53.1	11.52	0.27	11.51	0.01	3.59	0.24	3.62	-0.03
	1986	58.1	11.92	0.27	11.98	-0.06	4.14	0.19	3.98	0.16
rms				0.19		0.05		0.26		0.15

Table 3.1. *Absolute coordinates of the DNA fragments in the 2D display. The table indicates the size of each fragment, its GC content, and, for each dimension, the experimental absolute position (y_{exp}), the experimental uncertainty (σ_{exp}, in cm), the calculated absolute position (y_{calc}) and the error ($\Delta y = y_{exp} - y_{calc}$). Absolute positions in the first dimension were obtained with the expression of mobility in equation (3.4) and parameters* $\mu_L = 0.17 \times 10^{-4}$ *cm²/(V s),* $\mu_S = 4.53 \times 10^{-4}$ *cm²/(V s) and* $m = 41200$. *Absolute positions in the second dimension were obtained with the expression of mobility in equation (3.6), the expression of equivalent temperature in equation (3.8), and parameters* $L_r = 100$ *bps,* $[\mathrm{Na}^+] = 0.134$ *M,* $T_0 = 60$ *°C and* $\alpha = 0.540$ *°C.*

As will be seen in more detail in section 3.4, the calculation of $\mu(s, y(t))$ during DGGE requires the estimation of the number of open base pairs of sequence *s* at a temperature *T*, which has the same effect as the local concentration of denaturant. The modeling of denaturation was achieved by using the open source program MeltSim [38]. I used the set of thermodynamic parameters of Blake and Delcourt [39] and set the positional map resolution to 1, which corresponds to the highest possible calculation precision for the number of open base pairs. The influence of the remaining free parameter of the program, namely the salt concentration [Na^+], will be discussed in detail in section 3.4.

At last, the mobility $\mu(s, y)$ is expressed for both electrophoresis steps as a function of a certain numbers of parameters, which need to be adjusted to reproduce experimental results accurately. Therefore, I embedded equation (3.2) and the step by step integration of equation (3.3) in a refinement loop based on the gradient method, in order to vary the parameters so as to minimize the root mean square deviation between experimental positions and those calculated from equations (3.2) and (3.3).

3.3. Separation according to size

Van Winkle, Beheshti and Rill (vWBR) [75,76] recently proposed an empirical formula that correctly reproduces the observed mobilities of DNA fragments for a large number of experimental conditions. This formula writes

$$\frac{1}{\mu(s)} = \frac{1}{\mu_L} - \left(\frac{1}{\mu_L} - \frac{1}{\mu_S} \right) \exp\left(\frac{-N(s)}{m} \right) \tag{3.4}$$

where μ_L and μ_S are the respective mobilities of infinitely large and very small fragments, $N(s)$ is the length of the investigated DNA fragment, and *m* denotes the typical size that separates "small" from "large" sequences. Van Winkle *et al* [76] furthermore published the following expressions for μ_L, μ_S and *m* :

$$\begin{aligned} \mu_L &= 1.99 \times 10^{-4} \exp(-1.59\, C) \\ \mu_S &= (3.56 - 0.58\, C) \times 10^{-4} \\ m &= 7490 + 2780\, C \end{aligned} \tag{3.5}$$

where μ_L and μ_S are expressed in $cm^2/(V\ s)$ and *m* in base pairs, while *C* denotes the agarose gel concentration in percents. The six numerical constants in equation (3.5), as well as the forms of the equations themselves, are expected to be valid only for the precise system investigated by van Winkle *et al* [76]. Still, the somewhat different experimental conditions of reference [3] could be accounted for by feeding in equation (3.5) an adjusted gel concentration *C*=0.75% close to the exact value *C*=0.80%. It was indeed shown that this leads to calculated relative positions in good agreement with observed ones [3] (note, however, that absolute positions display errors larger than 1 cm). Due to the rather rigid forms of equations (3.5), it is however not warranted that this kind of adjustment will prove to be sufficient for experimental conditions that differ more widely from those of reference [76], in particular for the very popular polyacrylamide gels, and the choice of the additional parameter(s) to adjust might become rather tricky.

I found that a very efficient alternative to bypass this numerical problem consists in adjusting directly the parameters μ_L, μ_S and *m* of equation (3.4) against the final absolute locations along the first dimension. I obtained $\mu_L = (0.17 \mp 0.02) \times 10^{-4}$ $cm^2/(Vs)$, $\mu_S = (4.53 \mp 0.03) \times 10^{-4}$ $cm^2/(V\ s)$ and $m = 41200 \mp 6400$, which differs substantially from the values derived from equation (3.5) with the adjusted gel concentration *C*=0.75%, namely $\mu_L = 0.60 \times 10^{-4}$ $cm^2/(V\ s)$, $\mu_S = 3.12 \times 10^{-4}$ $cm^2/(V\ s)$ and $m = 9575$. Absolute positions obtained from equation (3.4) and the adjusted values of μ_L, μ_S and *m* are compared to observed ones in table 3.1. Experimental positions correspond to the average of the coordinates measured in three different experiments, while the associated uncertainties were estimated by taking the standard deviations for these three experiments. Note that the results of a fourth experiment, which differ markedly from the three other ones, were discarded. It can be seen that the root mean square deviation between observed and calculated absolute positions, that is 0.05 cm, is almost four times smaller than the average experimental uncertainty, which is 0.19 cm.

3.4. Separation according to sequence composition

It appears that very few studies have addressed the question of the electrophoretic mobility of partially melted DNA sequences. To my knowledge, there is indeed only one available model [68], which is inspired from previously existing results for the mobility of branched polymers in gels. Although this model has no firm theoretical background and should be tested under a larger range of experimental conditions, several studies performed so far have reported fairly good agreement with experimental data [77,78]. According to this model, the mobility of a partially melted DNA sequence decreases exponentially with the size of the melted regions, that is

$$\mu(s,T) = \mu_0(s)\exp\left(-\frac{p(T)}{L_r}\right) \tag{3.6}$$

where $\mu_0(s)$ is the mobility of the fragment when it is completely double-stranded, $p(T)$ is the sum of the probabilities for each base pair to be open at temperature *T*, and L_r is a size parameter, which is related to the mechanism that slows down partially melted fragments, and is therefore expected to depend on gel properties (concentration and pore size) and the flexibility of single-stranded DNA. Values of L_r reported in the literature range from 45 to 130 base pairs [77,78]. As already mentioned, I used the open source program MeltSim [38], together with the set of thermodynamic parameters of Blake and Delcourt [39], to estimate $p(T)$. The input quantities of this program are the temperature *T*, but also the salt concentration [Na^+]: it is indeed well-known that the melting temperature of a sequence varies logarithmically with [Na^+]. It should however be stressed that MeltSim was developed to predict the denaturation behavior of DNA sequences in cells and closely related media. Since porous gels differ sensitively from such solutions, it is not obvious that salt concentration has the same effect in cells and in gels. Moreover, it is difficult to predict how the presence of other salts in the composition of the buffer affects the melting temperature of the DNA sequences. In the simulations reported below, I therefore considered the [Na^+] input of the MeltSim program as a free parameter not necessarily related to the exact salt concentration in the gel.

Equation (3.6) is sufficient to calculate the mobility of DNA fragments in TGGE experiments, that is, when a temperature gradient is imposed to the gel, because the temperature *T* at each position *y* of the gel is known up to a certain precision. The link between the mobility $\mu(s, y)$ of equation (3.1) and the mobility $\mu(s, T)$ of equation (3.6) is therefore straightforward. This is no longer the case for DGGE experiments, where the temperature of the plate is kept uniform around 60°C and a gradient of chemical denaturant (urea+formamide) is added to the gel in order to destabilize base pairings. In this later case, the known quantity is the concentration C_d of the denaturant at each position *y* in the gel, so that estimation of the mobility $\mu(s, y)$ requires the additional knowledge of the equivalent temperature *T*, which has the same effect as a denaturant concentration C_d from the point of view of the melting of DNA fragments. A linear relation was proposed in reference [69], namely

$$T = 57 + \frac{1}{3.2} C_d \tag{3.7}$$

where C_d is the concentration of the standard stock solution of urea and formamide at position *y* (expressed in % v/v) and *T* the equivalent temperature (expressed in °C) to feed in the MeltSim program to estimate $p(T)$ at this position. As will be discussed below, equation (3.7) does not enable one to reproduce the absolute positions reported in reference [3]. I have therefore replaced equation (3.7) by the more general linear relation:

$$T = T_0 + \alpha C_d \tag{3.8}$$

where T_0 and α are considered as free parameters. I also took into account the very slight increase in gel viscosity due to the gradient of denaturant by slightly adjusting the mobilities of the DNA sequences at each time step. This is done by dividing the calculated mobility at each moment by the relative viscosity, which is computed according to reference [69]:

$$\eta_{rel} = 1 + 4.3 \times 10^{-3} C_d \tag{3.9}$$

To summarize, calculation of the mobility of DNA fragments in the second dimension requires the knowledge of the numerical values of four constants, namely [Na^+], L_r, T_0 and α. To be really complete, one should actually include $\mu_0(s)$, the mobility of fragment *s* when it is completely double-stranded (see equation (3.6)), in the list of the free parameters of the model. However, several trials showed that this parameter is so strongly correlated to the four other ones

that it is numerically impossible to let all of them vary simultaneously. I have therefore considered that the mobility $\mu_0(s)$ that appears in equation (3.6) is equal to the mobility obtained from equation (3.4) (in the first gel). This, of course, involves some degree of approximation, since the gels in the two dimensions are not identical.

I varied [Na^+], L_r, T_0 and α in order to reproduce the experimental results of reference [3]. These DGGE experiments were performed with 9 cm long plates and a denaturant concentration C_d increasing regularly from 25% to 100% between the extremities of the plates (the total concentration of stock denaturant was computed using the protocol given by Myers *et al* [79]: 100% stock denaturant corresponds to 7 M urea and 40% deionized formamide). Similarly to the first dimension, the experimental absolute positions and uncertainties reported in table 3.1 were obtained from three different experiments, while the results of a fourth experiment, which differ markedly from those of the three other ones, were discarded. I first allowed the four parameters to vary simultaneously. This resulted in the salinity [Na^+] decreasing below 0.001 M, which is the limit of validity of the set of thermodynamic parameters in the MeltSim program. In order to understand why this happens, I next performed a series of three parameters fits (L_r, T_0 and α) at several values of [Na^+] ranging from 0.001 M to 0.3 M. Results are shown in figures 3.3 and 3.4. It is seen in figure 3.3 that the root mean square (rms) error between experimental and calculated absolute positions actually remains essentially constant in the whole range 0.001-0.3 M. Furthermore, examination of figure 3.4 indicates that the adjusted values of T_0 and α vary logarithmically with [Na^+]. This is not really surprising because, as I have already mentioned, the melting temperature of a given sequence increases logarithmically with [Na^+]. At last, it is seen in the top plot of figure 3.4 that the adjusted value of L_r varies between 100 and 140 base pairs, which agrees with previously reported values [77,78].

Figures 3.3 and 3.4 are however not sufficient to illustrate how broad the space of solutions is, that is, how widely each parameter can be varied while still preserving a very good agreement between observed and calculated absolute positions. To get a better insight, figure 3.5 shows the results of a series of two parameters fits, which consisted in adjusting simultaneously T_0 and α for increasing values of L_r at two fixed values of [Na^+], namely 0.01 and 0.1 M. It is seen in the top plot of figure 3.5 that L_r can actually be varied between 30 and 220 base pairs without letting the rms error increase by more than 0.05 cm. As shown in the middle and bottom

plots of figure 3.5, the adjusted values of T_0 and α vary little with L_r in this range and remain close to $T_0 = 37\,°C$ and $\alpha = 0.63\,°C$ at [Na^+]=0.01 M, and $T_0 = 58\,°C$ and $\alpha = 0.54\,°C$ at [Na^+]=0.1 M. It should be clear from the examination of figures 3.3-3.5 that the numerical criterion alone is not sufficient to fix unambiguously the set of parameters to use in the model and that other criteria must be taken into account. In my opinion, a very sensible criterion would consist in fulfilling the condition that the equivalent temperature deduced from equation (3.8) should be equal to the true temperature of the plate in the absence of chemical denaturant, that is for $C_d = 0\,\%$. This amounts to impose $T_0 = 60\,°C$ in equation (3.8).

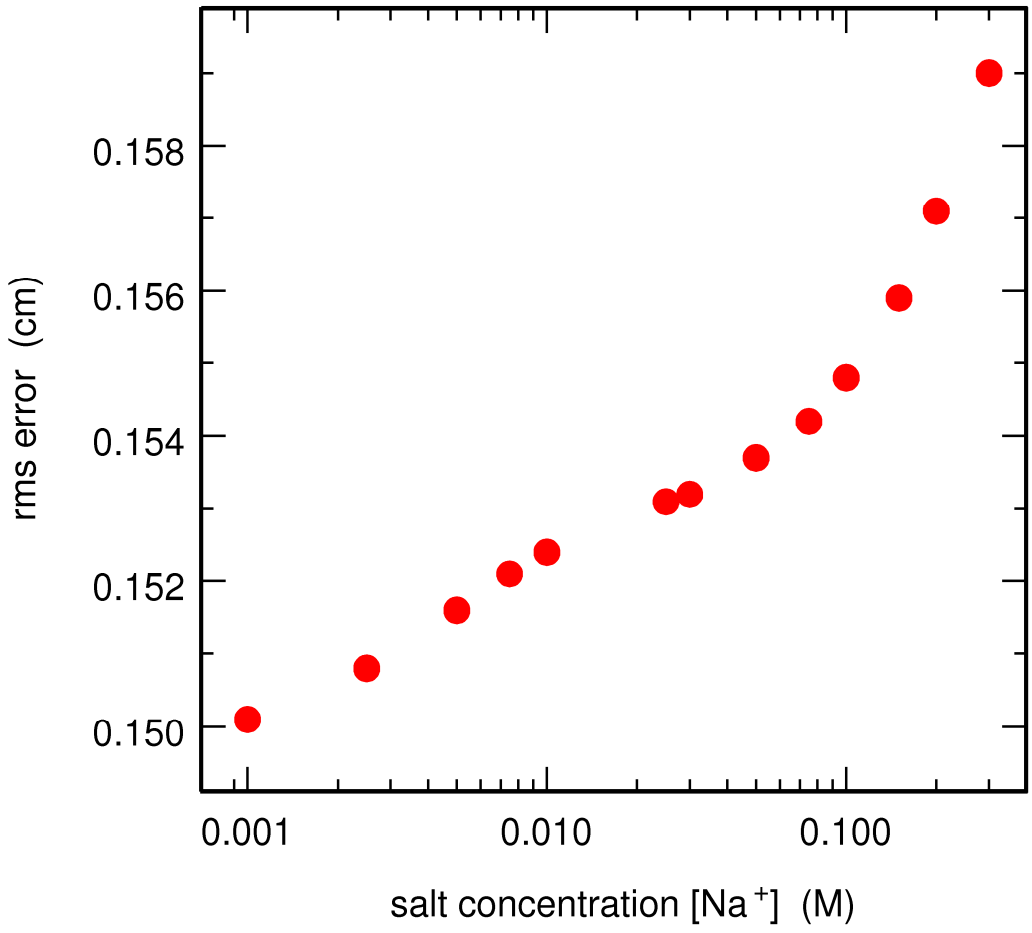

Figure 3.3. *Root mean square deviations between experimental and calculated absolute positions along the second dimension in the DGGE experiments for the 40 DNA sequences listed in table 3.1. The three parameters, L_r, T_0 and α were adjusted simultaneously for each fixed value of the salinity [Na^+].*

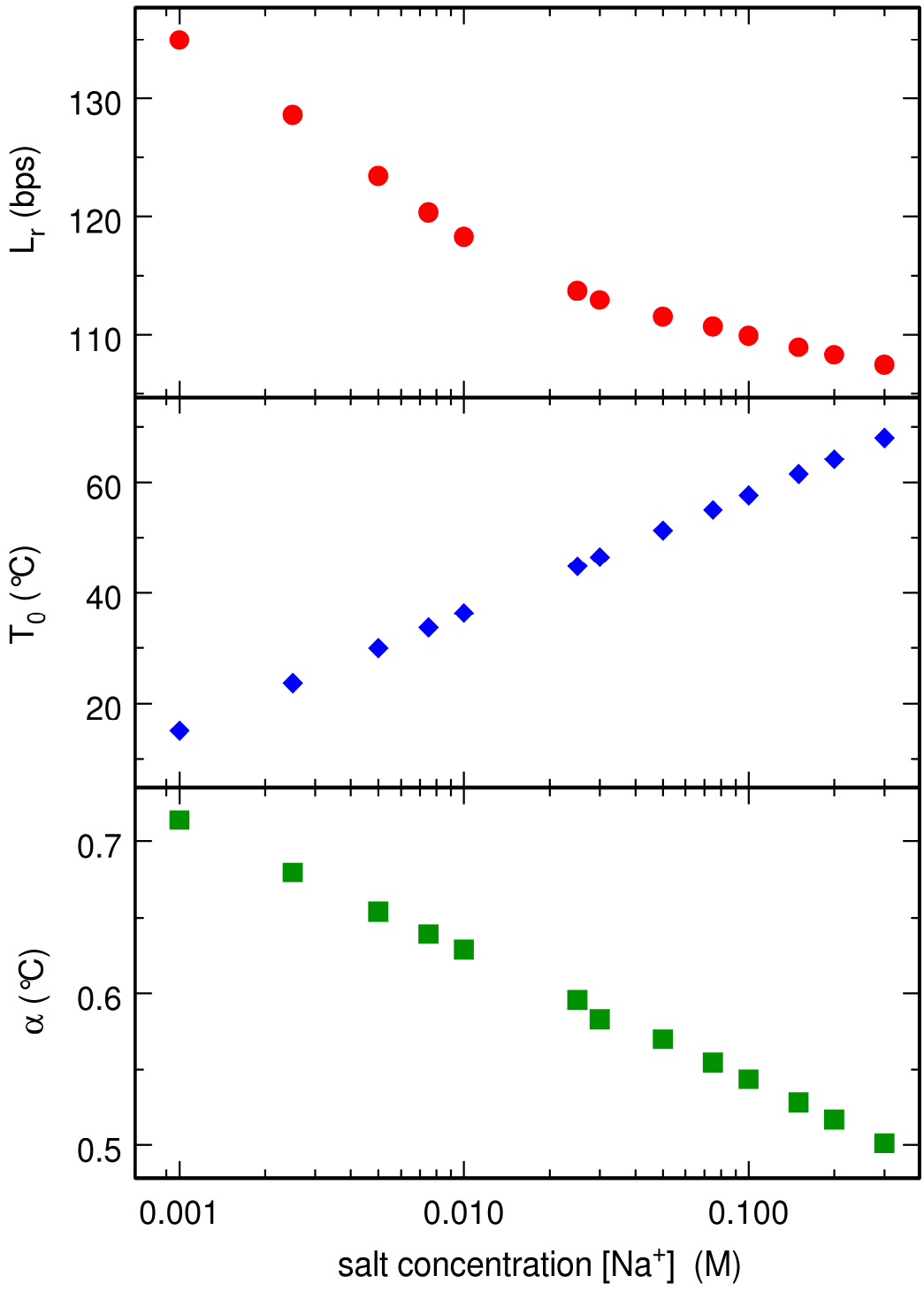

Figure 3.4. *Adjusted values of* L_r *(top plot),* T_0 *(middle plot), and* α *(bottom plot), for fixed values of the salinity* [Na^+].

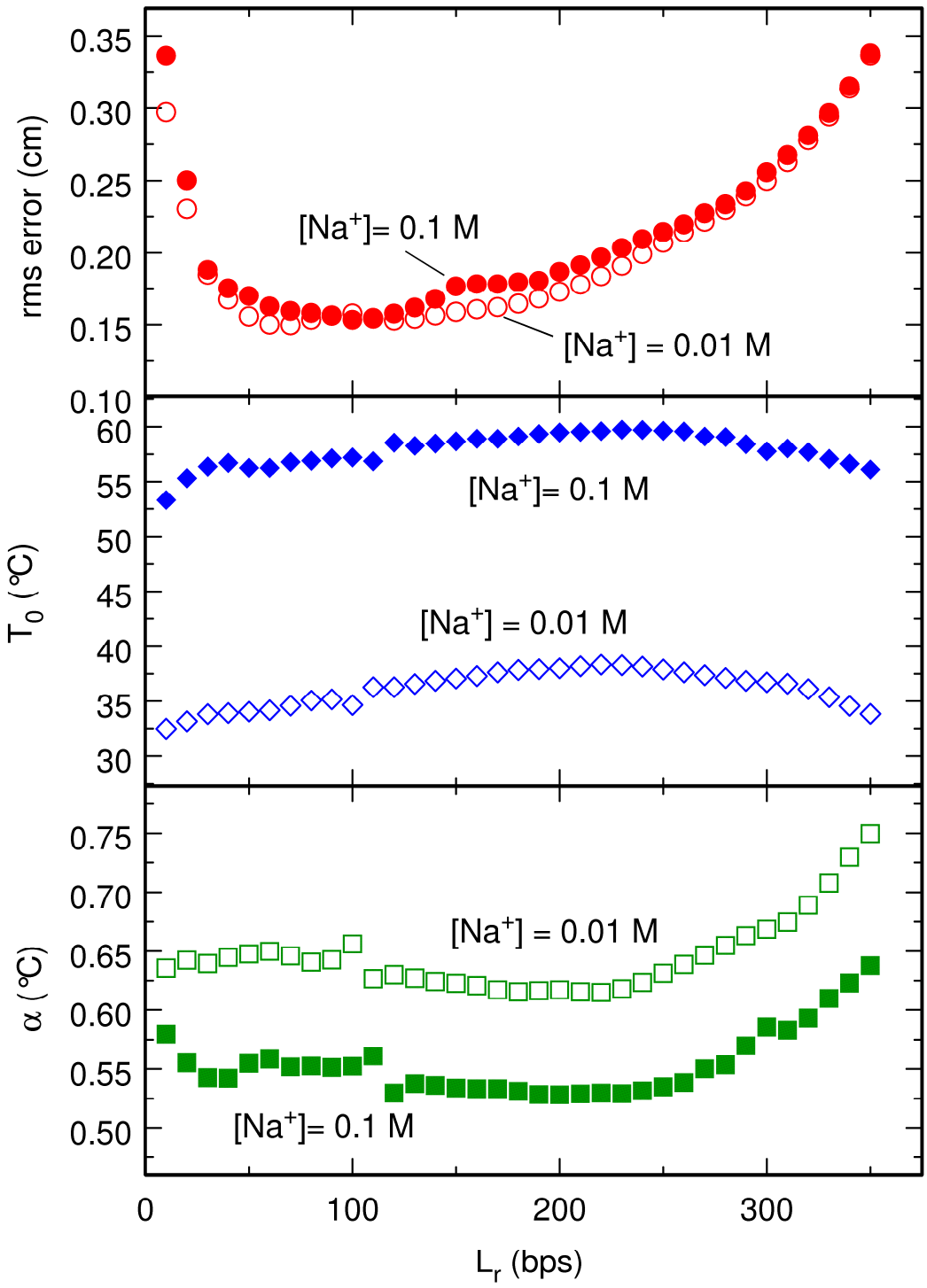

Figure 3.5. *Results of a series of two parameters fits, which consisted in adjusting simultaneously* T_0 *and* α *for increasing values of* L_r *at two fixed values of [Na+] (0.01 and 0.1 M). The top plot shows the root mean square error (expressed in cm) between experimental and calculated absolute positions along the second dimension of the DGGE experiments for the 40 DNA sequences listed in table 3.1. The middle plot shows the evolution of* T_0 *(expressed in °C) and the bottom plot the evolution of* α *(expressed in °C).*

I therefore performed another series of two parameters fits, which consisted in adjusting simultaneously [Na^+] and α for increasing values of L_r at fixed $T_0 = 60\,°C$. Results are shown in figure 3.6. Not surprisingly, the top plot again indicates that L_r can be varied between 30 and 220 base pairs without letting the root mean square error increase by more than 0.05 cm. What is, however, more interesting, is that the middle and bottom plots of figure 3.6 show that the value of [Na^+] to feed in the MeltSim program must be chosen in the range 0.10 to 0.15 M and that α consequently varies in the range 0.52 to 0.55 °C. Note that this is substantially larger than the value α=1/3.2=0.31 °C proposed in reference [69], but figure 3.6 unambiguously indicates that the absolute positions measured in reference [3] cannot be reproduced with such a low value of α - at least as long as one considers that $\mu_0(s)$ in equation (3.6) is equal to the mobility in the first dimension obtained from equation (3.4).

A second criterion is clearly mandatory in order to choose between the various solutions shown in figure 3.6. In my opinion, this criterion should rely on the knowledge of the number of base pairs of each sequence that are open at the end of the electrophoresis experiment. It should indeed be realized that all the solutions shown in figure 3.6 lead to the same dynamics of the fragments, that is, the mobility and the final position of each fragment do not depend on the chosen $(L_r, [Na^+], \alpha)$ triplet, but they do *not* lead to the same denaturation properties, that is, to the same number of open base pairs. Stated in other words, $p(T)/L_r$ remains the same for all $(L_r, [Na^+], \alpha)$ triplets, but not $p(T)$. This is clearly illustrated in figure 3.7, which shows the evolution as a function of time of the number of open base pairs (top plot) and of the position *y* (middle plot), as well as the evolution as a function of *y* of the mobility $\mu(s, y)$ (bottom plot), for two fragments with respective low and high GC contents and two $(L_r, [Na^+], \alpha)$ triplets with very different values of L_r. More precisely, the two fragments are the 2323 base pairs Eco91I digest of the λ-phage with 57.8% GC content and the 6555 base pairs Eco47I digest of the λ-phage with 38.0% GC content, while the chosen sets of parameters are $L_r = 30$ base pairs, $[Na^+] = 0.145$ M and $\alpha = 0.526$ °C, and $L_r = 200$ base pairs, $[Na^+] = 0.105$ M and $\alpha = 0.524$ °C. Examination of the middle and bottom plots of figure 3.7 indicates that the calculated positions and mobilities of the two fragments are very similar for the two sets of parameters. In contrast, it can be seen in the top plot that the number of base pairs that are open at the end of the

electrophoresis experiment differ widely for the two sets of parameters: the set with $L_r = 30$ base pairs predicts that about 250 base pairs are open for both fragments, while the set with $L_r = 200$ base pairs predicts that this number is close to 1500 (note that $250/30 \approx 1500/200$). In order to fix unambiguously the correct set of parameters, which must be used to predict electrophoresis experiments such as those reported in reference [3], one should therefore complement these experiments with detailed measurements of the mobility of a few sequences, as in figure 4 of reference [77]. The positions of the bumps in the evolution of mobility, which reflect the abrupt opening of large portions of the fragment, indeed reveal the correct value of L_r, and consequently also of $[Na^+]$ and α.

Since these additional data are not available for the experiments reported in reference [3], we chose the set of parameters that leads to the smallest root mean square error, that is $L_r = 100$ base pairs, $[Na^+] = 0.134$ M and $\alpha = 0.540$ °C, to compare calculated and experimental absolute positions in the second dimension. Results are tabulated in the four last columns of table 3.1. It is stressed that the root mean square deviation between calculated and observed absolute positions (0.15 cm) is almost twice smaller than the average experimental uncertainty (0.26 cm).

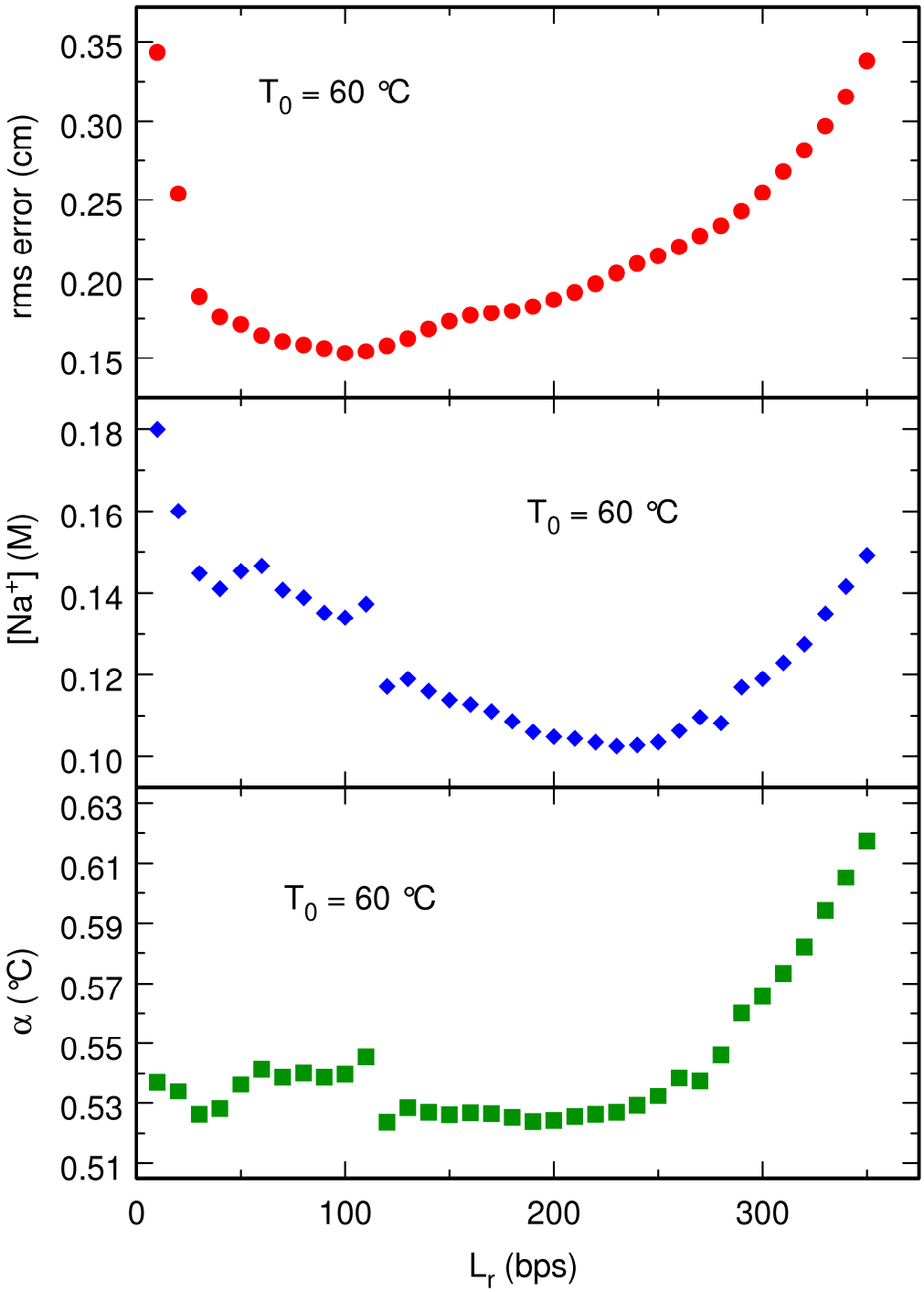

Figure 3.6. *Results of a series of two parameters fits, which consisted in adjusting simultaneously* $[Na^+]$ *and* α *for increasing values of* L_r *at fixed* $T_0 = 60\,°C$. *The top plot shows the root mean square error (expressed in cm) between experimental and calculated absolute positions along the second dimension (DGGE experiments) for the 40 DNA sequences listed in table 3.1. The middle plot shows the evolution of* $[Na^+]$ *(expressed in M) and the bottom plot the evolution of* α *(expressed in °C).*

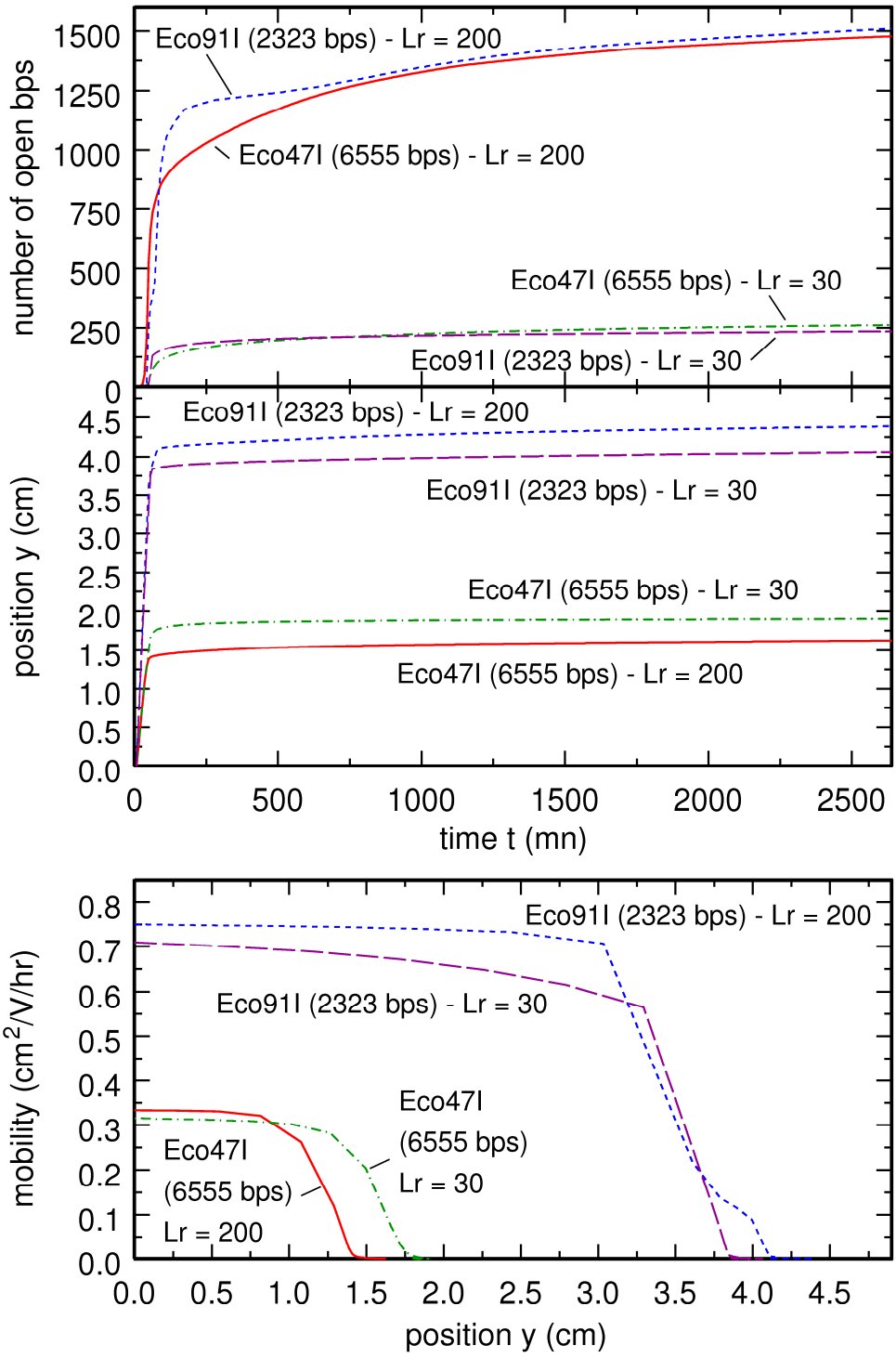

Figure 3.7*. Evolution as a function of time of the number of open base pairs (top plot) and of the position y (middle plot), and evolution as a function of y of the mobility* $\mu(s,y)$ *(bottom plot), for two different fragments and two different sets of parameters. The two fragments are the 2323 bps Eco91I digest with 57.8% GC content and the 6555 bps Eco47I digest with 38.0% GC content. The two sets of parameters are* $L_r = 30$ *bps,* $[\mathrm{Na}^+] = 0.145$ *M and* $\alpha = 0.526$ *°C, and* $L_r = 200$ *bps,* $[\mathrm{Na}^+] = 0.105$ *M and* $\alpha = 0.524$ *°C.* $T_0 = 60$ *°C for both sets.*

3.5. Conclusion

In this chapter, I have presented a study of the parameterization issues associated with a model aimed at predicting the final absolute locations of DNA fragments in 2D display experiments. In particular, I have shown that simple expressions for the mobility of DNA fragments in both dimensions allow one to reproduce experimental final absolute locations to better than experimental uncertainties. I have furthermore pointed out that the results of 2D display experiments are not sufficient to determine the best set of parameters for the modeling of fragments separation in the second dimension, and that additional detailed measurements of the mobility of a few sequences are necessary to achieve this goal.

To model electrophoresis along the second dimension, which involves the melting of DNA along a concentration gradient of chemical denaturant, I have written a program that embeds the MeltSim code, which is based on the Poland-Scheraga model. This work convinced me that programs like MeltSim are very convenient for practical purposes. They are simple to use, even for people like me, who have not enough time at disposition to understand all the subtleties of the description of DNA melting with such statistical models. In addition, the precision of the underlying model is sufficient for many practical purposes. For example, it was mentioned in the discussion at the end of reference [3], that the weakest part of the simulation program is probably equation (3.6), which expresses the mobility of a partially melted DNA sequence as an exponentially decreasing function of the size of the melted regions, and that the L_r parameter should probably include some dependence on the properties of the gel (like its concentration and the size of the pores) and the studied DNA sequences (their length, the number and location of melted regions). All the attempts we made in this direction were unsuccessful, which is probably due to the fact that experimental uncertainties, which result essentially from the difficulty to control precisely the reproducibility of experimental conditions and the spontaneous deformation of the gels, are almost twice as large as the rms deviation between experimental and calculated positions. Today's limiting step is therefore neither the MeltSim code, nor the expressions I used to calculate mobility in both dimensions, but rather the experimental procedure: to my mind, it will indeed not be possible to improve the model as long as experimental uncertainties will not have been made substantially smaller than what can be achieved in today’s experiments.

4. Improving our dynamical DNA model

4.1. Introduction

It has been shown in the preceding chapter that statistical models enable a fast and reliable estimation of the denaturation properties of DNA sequences. This is no longer the case for the dynamical models, which will be discussed in this chapter. More precisely, dynamical models are indeed accurate - and we work hard to make them even more accurate - but they involve calculations, which are orders of magnitude more time consuming than statistical ones. As a consequence, they would for instance have been of no practical help to simulate the 2D electrophoresis experiments discussed in the previous chapter.

Yet, dynamical models are useful, because they provide a complementary point of view on the dynamics of melting, in the sense that, for dynamical models, the macroscopic properties of the sequence (like the critical temperature and the temperature evolution of the specific heat) depend only on its microscopic properties, like the depth and the shape of the stacking and pairing interactions. In contrast, models like that of Poland and Scheraga make heavy use of the statistical properties of the sequence, like the partition function of the loops and the cooperativity parameter. Moreover, the effect of temperature is explicitly plugged in statistical models through the definition of site-dependent stacking entropies. It is therefore interesting to check the degree of agreement between predictions obtained from models that rely on so different building blocks. I will come back to this point in section 4.5. Obviously, dynamical models are also powerful tools when one is not interested in the mean value of a quantity but rather in its fluctuations [40].

The purpose of the work presented in this chapter is threefold. I will first show how it is possible to get better estimations of the parameters of the model developed in our group by taking into consideration experimental facts that were disregarded up to this point. I will then compare the results obtained with the improved model with those obtained from statistical models. Finally, I will describe briefly the critical properties of the new model and compare them to those of the previous one. I will conclude with a discussion of the critical behaviour of the model in the very narrow region just below the critical temperature.

4.2. Adjustment of the parameters

As already emphasized in section 2.5, the dynamical model developed in our group needs to be improved with respect to at least three points. First, this model predicts, like the DPB one, a much too large sensitivity of the critical temperature with respect to the length of the sequence, leading to unrealistically small melting temperatures for sequences with less than several hundred base pairs. Moreover, we performed some calculations to probe the mechanical unzipping of DNA, and found that our model predicts too small critical forces. Additionally, the temperature resolution is too small compared to the experimental one. I will show in this section how these points can be improved, at least partially, by varying the parameters of the model.

The Hamiltonian *H* of the model, which we developed to study DNA denaturation, is shown in equation (2.11). This Hamiltonian describes free DNA, that is, the case where no external force is applied to the sequence. In the presence of a force acting on one of the bases of the base pair at position *n*=1, the energy $H_{stretch}$ of the perturbed system may be written as [80]:

$$H_{stretch} = H - F y_1 \tag{4.1}$$

It should however be noted that, because of the $\sqrt{2}$ factor that appears in the expression of y_n as a function of the positions u_n and v_n of each nucleotide (see equation (2.6)), *F* is, properly speaking, not the experimental force, but rather the experimental force multiplied by $\sqrt{2}$. The critical force $F_c(T)$ is defined as the force required to keep the two strands separated at temperature *T*. This is the force, for which the variation of the average free energy per base pair of a very long sequence, $g_0(T)$, is equal to the variation of free energy per base pair of the stretched single strands, $g_u(T,F)$ [60-61,81-82]:

$$g_u(T, F_c(T)) - g_u(T_c, 0) = g_0(T) - g_0(T_c) \tag{4.2}$$

Equation (4.2) contains an approximation, in the sense that it assumes that the free energy per base pair of a stretched double-stranded sequence is equal to that of an unstretched sequence. Results obtained using equation (4.2) are therefore better checked with independent calculations. Following the arguments of Singh and Singh [80], the variation of the free energy per base pair of unstretched long sequences may be estimated for the model in equation (2.11) according to

$$g_0(T) - g_0(T_c) = D\left(\frac{T}{T_c} - 1\right) \tag{4.3}$$

for temperatures T smaller than the critical temperature T_c. In equation (4.3), $D = D_n$ denotes the depth of the Morse potential for a homogeneous sequence. Moreover, it is possible for some models to calculate the free energy per base pair of the stretched single strands by taking the derivative of the partition function $Z(T,F)$

$$\begin{aligned} Z(T,F) &= \int \mathrm{d}y_1\, \mathrm{d}y_2 \ldots \mathrm{d}y_N \exp(-\beta H_{stretch}) \\ &= \int \mathrm{d}y_1\, \mathrm{d}y_2 \ldots \mathrm{d}y_N \exp\left(-\beta\left(N D + \sum_{n=2}^{N}\left\{W^{(n)}(y_n, y_{n-1}) + F(y_n - y_{n-1})\right\}\right)\right) \end{aligned} \tag{4.4}$$

$$g_u(T,F) = -\frac{1}{N\beta}\ln(Z(T,F))$$

where $\beta = 1/(k_B T)$. When approximating the nearest-neighbour interaction potential in equation (2.11) by

$$W^{(n)}(y_n, y_{n-1}) \approx \min\left[\frac{\Delta H}{C}, \frac{\Delta H\, b}{C}(y_n - y_{n-1})^2\right] + K_b(y_n - y_{n-1})^2 \tag{4.5}$$

one obtains

$$g_u(T,F) = D - \frac{1}{\beta}\ln(a(I_1 + I_2)) \tag{4.6}$$

where

$$\begin{aligned} I_1 &= \sqrt{\frac{\pi}{4\beta K_b}}\exp\left(u^2 - \frac{\beta\Delta H}{C}\right)\{2\text{-erf}(v\text{-}u) - \mathrm{erf}(v+u)\} \\ u &= \sqrt{\frac{\beta F^2}{4K_b}} \\ v &= \sqrt{\frac{\beta K_b}{b}} \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} I_2 &= \sqrt{\frac{\pi}{4\beta\kappa}}\exp(u'^2)\{\mathrm{erf}(v'\text{-}u') + \mathrm{erf}(v'+u')\} \\ \kappa &= K_b + \frac{\Delta H\, b}{C} \end{aligned}$$

$$u' = \sqrt{\frac{\beta F^2}{4\kappa}}$$
$$v' = \sqrt{\frac{\beta\kappa}{b}} \qquad (4.8)$$

The plot of $F_c(T)$ obtained with equations (4.2), (4.3) and (4.6) and the original set of parameters reported in section 2.5 is shown as a solid line in figure 4.1. It was checked that Monte-Carlo simulations performed with the Hamiltonian in equation (4.1), reported as open circles in figure 4.1, are in excellent agreement with this curve in the 290-370 K temperature range. Conclusion therefore is that the parameters used up to now lead to a too small critical force around 20°C (the experimental value lies in the range 10-20 pN [60,61]), especially when remembering that *F* in equations (4.1) to (4.8) denotes the experimental force multiplied by $\sqrt{2}$. Examination of equations (4.6) to (4.8) shows that $F_c(T)$ depends strongly on K_b, which was fixed somewhat arbitrarily in the original set of parameters. Comparison with mechanical unzipping experiments will therefore help fix this parameter to a more grounded value. Plots of $g_u(T,F)$ obtained from equation (4.6) indicate that K_b must actually be increased for the calculated critical curve to come closer to the experimental one. This is a very positive point, because we also noticed that K_b =10^{-5} eV Å^{-2} sometimes leads to distances between successive bases on the same strand that are unrealistically large. Increasing K_b will therefore improve the quality of the model with respect to two points and not only one.

Moreover, and despite the fact that we have no definitive proof thereof, many trials convinced us that the only way to substantially reduce the dependence of the critical temperature on the length of the sequence consists in decreasing the depth of the stacking interaction, that is, in assuming smaller values for $\Delta H_n / C$. Since we want to go on using the stacking enthalpies ΔH_n that were adjusted for statistical models, *C* must consequently be made larger than the value $C = 2$, which we used up to now and was obtained by assuming that paired bases do not unstack simultaneously [7]. Low frequency Raman spectra [84,85] and theoretical investigations of collective modes in DNA [86,87] suggest, on the other hand, that the stacking stiffness $\Delta H_n b / C$ may be larger than what was assumed in the original set of parameters (note, however, that the two models are substantially different). The increase of the parameter *b* controlling the width of the Gaussian hole must therefore be larger than the decrease of the hole depth $\Delta H_n / C$.

4.3. New parameters

Taking into account the considerations made above and requiring that calculated melting curves reproduce experimental ones is still not sufficient to fix all the parameters of the model unambiguously. However, we found that the following set of parameters allows for a correct reproduction of most of the known properties of DNA under usual salinity conditions (75 mM NaCl):

- Morse pairing potential: $D_n = D = 0.048$ eV and $a = 6.0$ Å^{-1} (against $D = 0.040$ eV and $a = 4.45$ Å^{-1} previously).
- Stacking potential: $C = 4$ and $b = 0.80$ Å^{-2} (against $C = 2$ and $b = 0.10$ Å^{-2} previously). For inhomogeneous sequences, the ten stacking enthalpies ΔH_n are taken from table 1 of reference [38] (as previously), while we used $\Delta H_n = \Delta H = 0.409$ eV for homogeneous ones (against the value $\Delta H = 0.44$ eV that was used previously).
- Backbone stiffness: K_b=4.0 10^{-4} eV Å^{-2} (against K_b=10^{-5} eV Å^{-2} previously). As anticipated, the most important change compared to the earlier set of parameters concerns this parameter.

The plot of $F_c(T)$ obtained with this new set of parameters and equations (4.2), (4.3) and (4.6) is shown in figure 4.1. It is seen that, although probably still somewhat too small, the critical force around 20°C is now in much better agreement with the experimentally determined one [60,61]. Moreover, the melting temperature also decreases much less rapidly as a function of the length of the sequence. Figure 4.2 indeed indicates a $540/N^{0.77}$ dependence, which is still larger than the decrease predicted by statistical models but is in qualitative agreement with both the $500/N$ dependence that is plugged in most online oligonucleotide property calculators and the experimental results reported in [59].

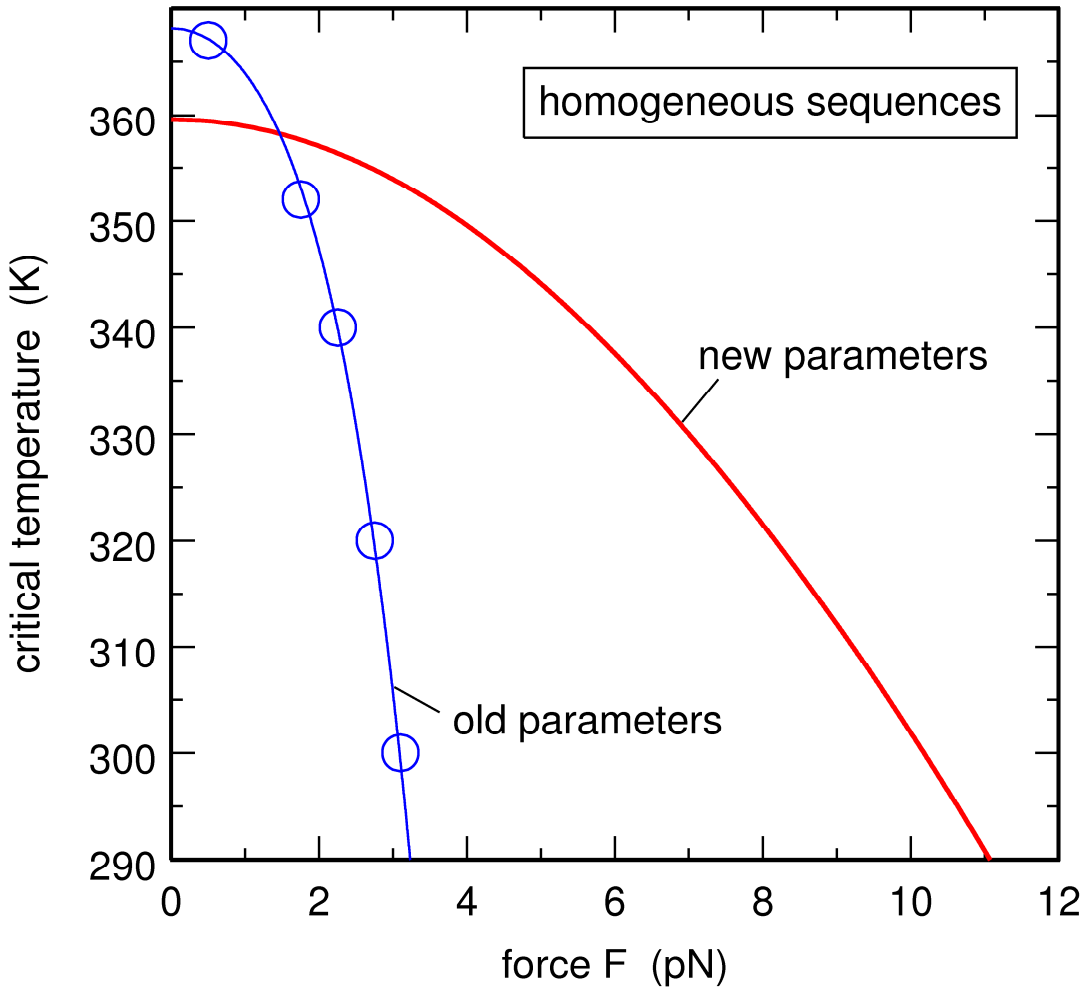

Figure 4.1. *Plot of the critical force F_c, which is required to keep the two DNA strands separated, as a function of the temperature T of the sequence, according to the model of equation (2.11) and the old and new sets of parameters. Solid lines were obtained from equations (4.2), (4.3) and (4.6) and the few open circles from Monte Carlo simulations, as a check to the validity of these equations.*

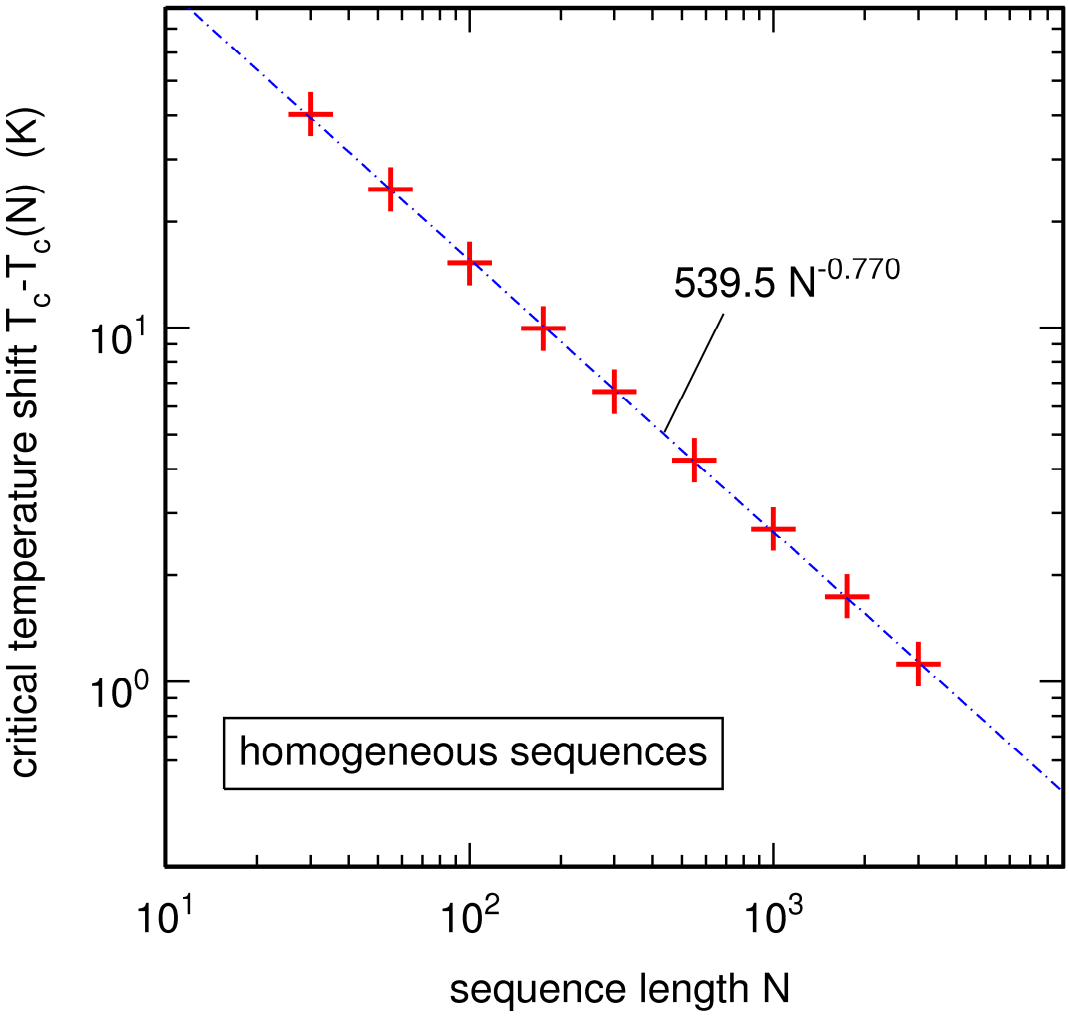

Figure 4.2. *Plot, as a function of the length of the sequence N, of the difference* $T_c - T_c(N)$ *between the critical temperatures of an infinitely long homogeneous sequence,* $T_c = 359.43$ *K (see section 4.6), and a homogeneous sequence with N base pairs,* $T_c(N)$. *The dot-dashed line is the least-square fit to the calculated shifts.*

4.4. Heterogeneous pairing and salt concentration contributions

The set of parameters proposed above assumes that heterogeneity is carried by stacking interactions. One might instead assume that heterogeneity is carried by pairing interactions, as in heterogeneous versions of the DPB potential [51,52,54]. It is sufficient, for this purpose, to fix ΔH_n to its average value $\Delta H = 0.409$ eV and introduce two different values for the Morse potential depth D_n, namely one for AT base pairs and one for GC base pairs. One is thus led to the following set of parameters:

- Morse pairing potential: $D_n = 0.041$ eV for AT base pairs, $D_n = 0.054$ eV for GC base pairs, and $a = 6.0$ Å^{-1}.
- Stacking potential: $C = 4$, $b = 0.80$ Å^{-2} and $\Delta H_n = \Delta H = 0.409$ eV.
- Backbone stiffness: K_b =4.0 10^{-4} eV Å^{-2}.

Results obtained with this set of parameters are qualitatively and quantitatively similar to those obtained with the set of parameters proposed in section 4.3. Since recent work suggests that heterogeneity is carried by both pairing and stacking interactions [88-90], one could think about introducing both different D_n and ΔH_n values in the model. We however made no attempt in this direction because of the complexity of TI calculations for this kind of hybrid models [53].

At last, it should be noted that the influence of different salinity conditions on DNA melting can easily be taken into account in this particular form of our model by using $D_n = 0.041 + 0.006\ \mathrm{Log}([\mathrm{Na}^+]/[\mathrm{Na}^+]_0)$ for AT base pairs and $D_n = 0.054 + 0.004\ \mathrm{Log}([\mathrm{Na}^+]/[\mathrm{Na}^+]_0)$ for GC base pairs, where D_n is expressed in eV and $[\mathrm{Na}^+]_0 = 75$ mM. The variations of critical temperature with respect to salinity obtained with these expressions agree well with those predicted by statistical models.

4.5. Comparison of the melting curves obtained with this model and statistical ones

As mentioned before, in statistical models a base pair can only assume one of two states: "open" or "closed". There is no ambiguity when estimating, for example, the fraction of open base pairs as a function of temperature, or the temperature at which each base pair of a given sequence has probability 0.5 to be open. Such plots are shown as dashed lines in figures 4.3 and 4.4 for the 1793 base pair human β-actin cDNA sequence (NCB entry code NM_001101). Calculations were performed with the MeltSim program [38], the parameters of Blossey and Carlon [35], and salinity $[\mathrm{Na}^+]_0 = 75$ mM.

"Closed" and "open" are more ambiguous concepts in the case of dynamical models, which are expressed in terms of continuous coordinates y_n. For example, one might consider that the fraction of open base pairs is obtained by computing at each time *t* the fraction of base pairs for which y_n is larger than a given threshold y_{thresh} and in subsequently averaging this quantity over *t* [6,52,54,55]. Alternatively, one can consider that a given base pair *n* is open if the mean elongation $\langle y_n \rangle$ is larger than the threshold y_{thresh} and average this quantity over the sequence. The two definitions are rather close and, as long as one does not deal with experimental results obtained with ultra-short laser pulses, there is no physical reason to choose one definition instead of the other. Still, the curves obtained with these two definitions are not identical. In particular, we noticed that results obtained with the second definition are better resolved in temperature and closer to those obtained with statistical models [53,58]. Further on, we will therefore use this definition and consider that base pair *n* is open if $\langle y_n \rangle > y_{thresh}$. It remains that the choice of y_{thresh} itself is not trivial. Figure 2.4 indeed shows that if one chooses for y_{thresh} a too small value, like for example two or three times $1/a$ (approximately 0.5 Å), then application of the criterion to a long homogenous sequence would lead to the erroneous conclusion that all base pairs are already open tens of Kelvins below the critical temperature. For such long homogeneous sequences, the larger the value of y_{thresh}, the closer the critical temperature determined with the $\langle y_n \rangle > y_{thresh}$ criterion to the exact one. But, on the other hand, a too large value of y_{thresh} is in

turn not suitable for inhomogeneous sequences, because different portions of an inhomogeneous sequence melt at different temperatures and the separation of open base pairs belonging to bubbles is limited by the double-stranded portions. The choice of y_{thresh} therefore appears as a compromise between these two conflicting considerations. Practically, we found that, for the set of parameters proposed in section 4.3, the choice $y_{thresh} = 10$ Å leads to reasonable results for both homogeneous and inhomogeneous sequences. Still, one must keep in mind that the critical temperature determined with this criterion is 2 to 3 Kelvins lower than the correct one (see figure 2.4).

We computed the evolution of the fraction of open base pairs as a function of temperature and the melting temperature of each base pair of the 1793 base pairs actin sequence for the parameters of section 4.3 using the TI procedure described in reference [53]. The results are plotted as solid lines in figures 4.3 and 4.4. The results obtained with three different thresholds ($y_{thresh} = 7$, 10 and 15 Å) are shown for the sake of comparison. It can be checked that, except for the short region of the sequence that melts at the highest temperature, the agreement between results obtained with statistical and dynamical models is rather striking. In particular, the resolution in temperature of melting curves is higher for the new parameters than for the old set and almost comparable to that of statistical models. In contrast, no increase in resolution is observed in the plot of $c_V(T)$, as can be checked in figure 4.5.

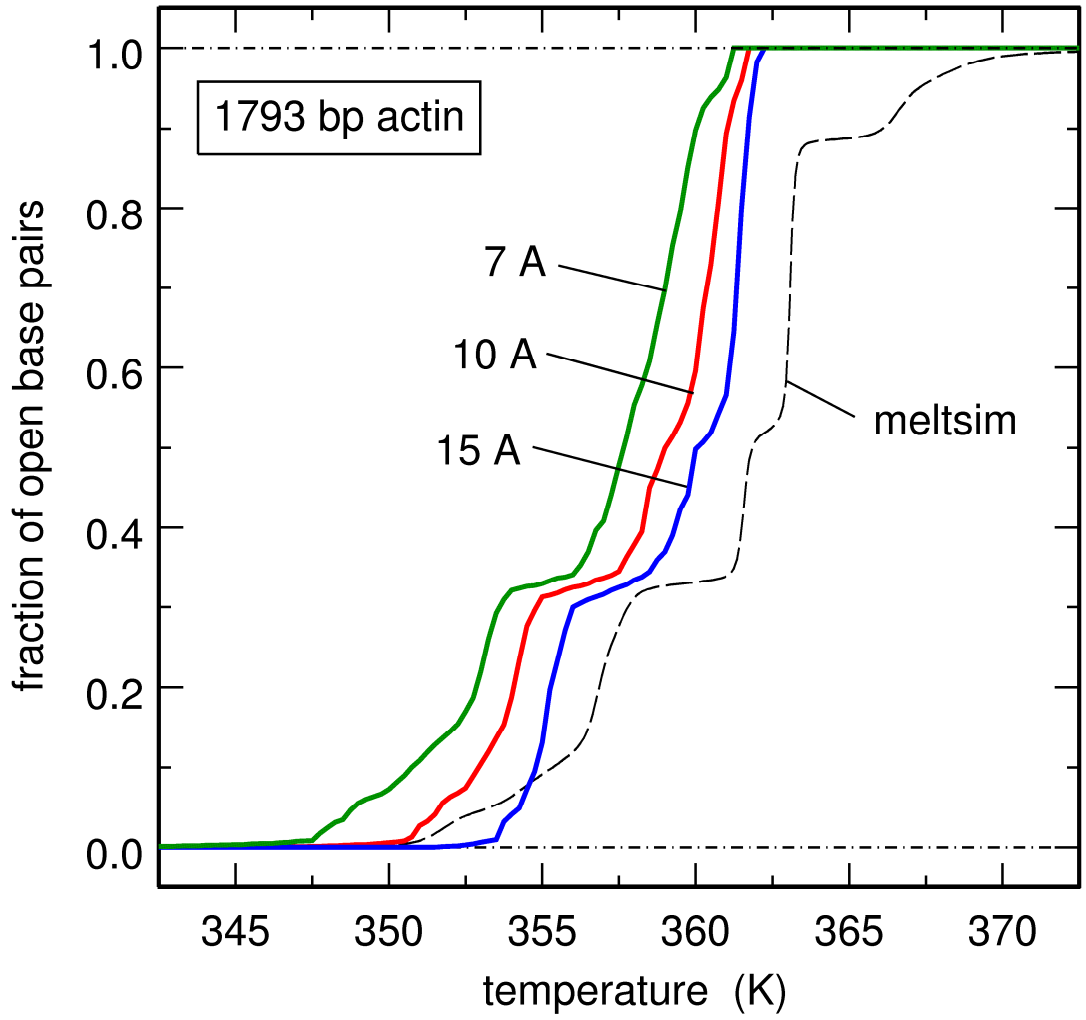

Figure 4.3. *Plot, as a function of the temperature T of the sequence, of the fraction of open base pairs for the 1793 base pairs human β-actin cDNA sequence (NCB entry code NM_001101) obtained with MeltSim [38] (dashed line) and the dynamical model proposed here (solid lines). MeltSim calculations were performed with the parameters of Blossey and Carlon [35] and a salt concentration* $[\mathrm{Na}^+]_0 = 75$ *mM. Results obtained with three different thresholds (* $y_{thresh} = 7$, *10 and 15 Å) are shown for TI calculations performed with the dynamical model. Remember that critical temperatures determined with the* $\langle y_n \rangle > y_{thresh}$ *criterion are 2 to 3 Kelvins lower than exact ones, as discussed in section 4.5.*

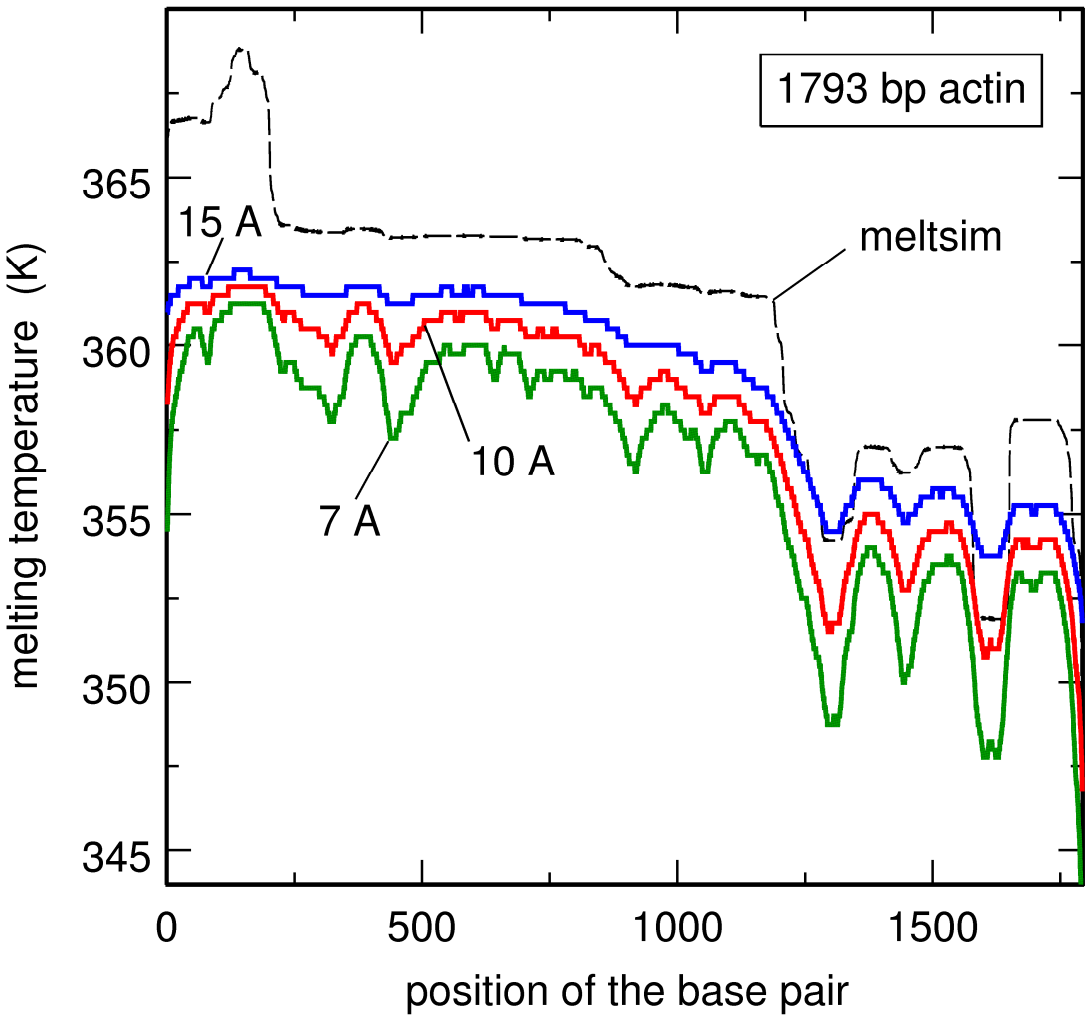

Figure 4.4. *Plot, as a function of the position of the base pair, of the opening temperature of each base pair of the 1793 base pairs human β-actin cDNA sequence (NCB entry code NM_001101) obtained with MeltSim [38] (dashed line) and the dynamical model proposed here (solid lines). MeltSim calculations were performed with the parameters of Blossey and Carlon [35] and a salt concentration* $[\mathrm{Na}^+]_0 = 75$ *mM. Results obtained with three different thresholds (* $y_{thresh} = 7$*, 10 and 15 Å) are shown for TI calculations performed with the dynamical model. Remember that critical temperatures determined with the* $\langle y_n \rangle > y_{thresh}$ *criterion are 2 to 3 Kelvins lower than exact ones, as discussed in section 4.5.*

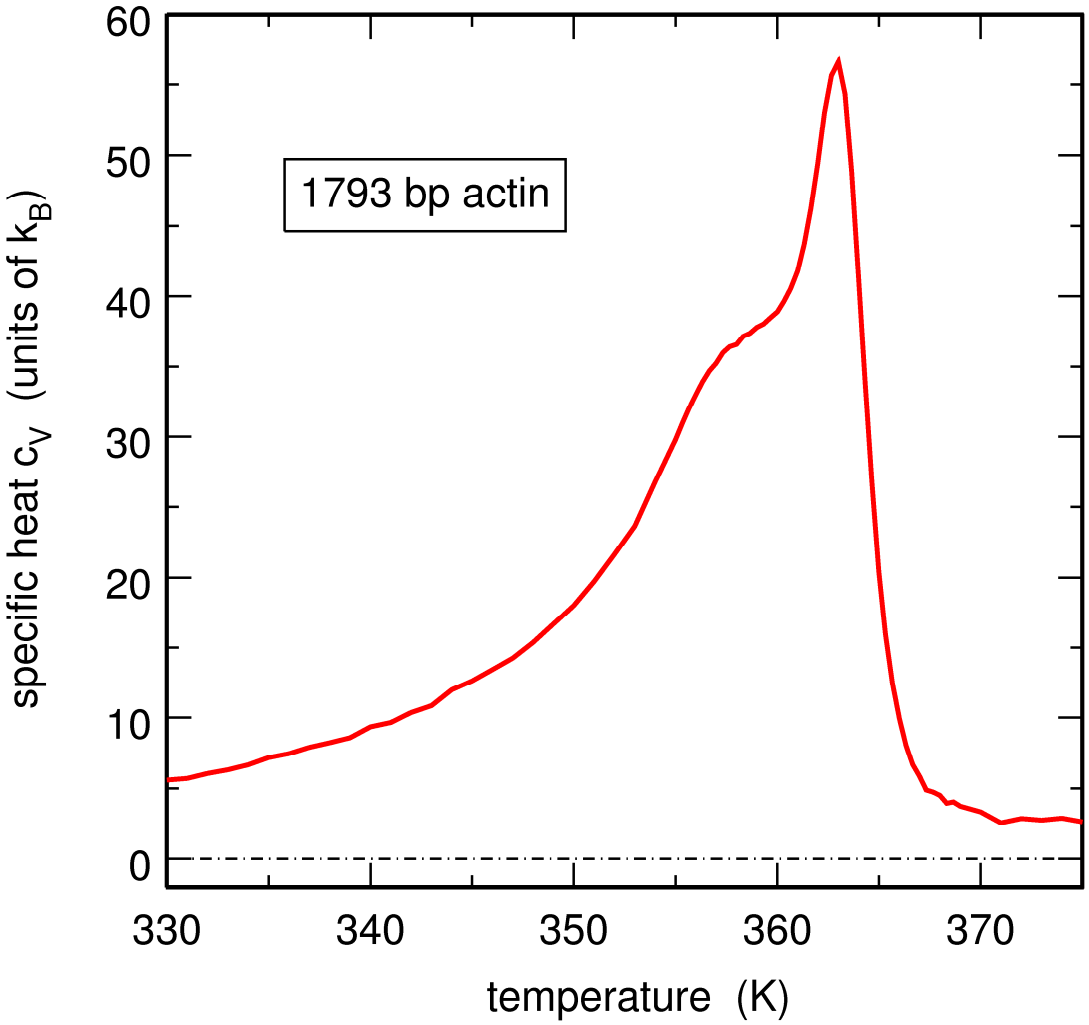

Figure 4.5. *Plot of the specific heat per particle, c_V, as a function of temperature T for the 1793 base pairs human β-actin cDNA sequence (NCB entry code NM_001101), obtained from TI calculations performed with the dynamical model proposed here. c_V is expressed in units of the Boltzmann constant k_B.*

4.6. Critical behaviour of the dynamical model

This section is devoted to the description of the critical behaviour of the dynamical model with the parameters proposed in section 4.3. I will show that it does not differ significantly from the behaviour observed with the set of parameters used in references [56,57], which implies that the melting of homogeneous DNA sequences looks like a first-order phase transition. It will be however pointed out that there is necessarily a crossover to another regime very close to the melting temperature.

The temperature evolution of the entropy per base pair, *s*, is shown in figure 4.6 for infinitely long sequences and sequences with *N*=1000 and *N*=100 base pairs. This plot, as well as all the other plots discussed in this section, were obtained from TI calculations performed as discussed in references [56,57]. It is seen that the temperature evolution of *s* displays the step-like behavior that is characteristic of first-order phase transitions. This is particularly clear for the infinite sequence and the sequence with *N*=1000 base pairs, but the step-like behavior is still well-marked for shorter sequences. As usual, this step-like behavior of *s* corresponds to thin peaks in the temperature evolution of the specific heat per base pair, c_V, as can be checked in figure 4.7. Note that, in both figures, the solid line associated with infinite sequences is interrupted in the narrow temperature interval where TI calculations are not valid.

Further information is gained by calculating the critical exponents, which characterize the power-law behavior of several statistical properties of infinitely long homogeneous sequences close to the critical temperature. For example, critical exponents α, β and ν are defined according to

$$\begin{aligned} c_V &\propto (T_c - T)^{-\alpha} \\ \langle y \rangle &\propto (T_c - T)^{\beta} \\ \xi &\propto (T_c - T)^{-\nu} \end{aligned} \qquad (4.10)$$

where ξ denotes the correlation length and $\langle y \rangle$ is taken as the order parameter of the melting transition [56]. The critical temperature of a sequence of length *N*, $T_c(N)$, is easily found as the temperature where c_V is maximum. Because of the temperature interval where TI calculations are not valid, it may be somewhat more complex to determine the critical temperature of

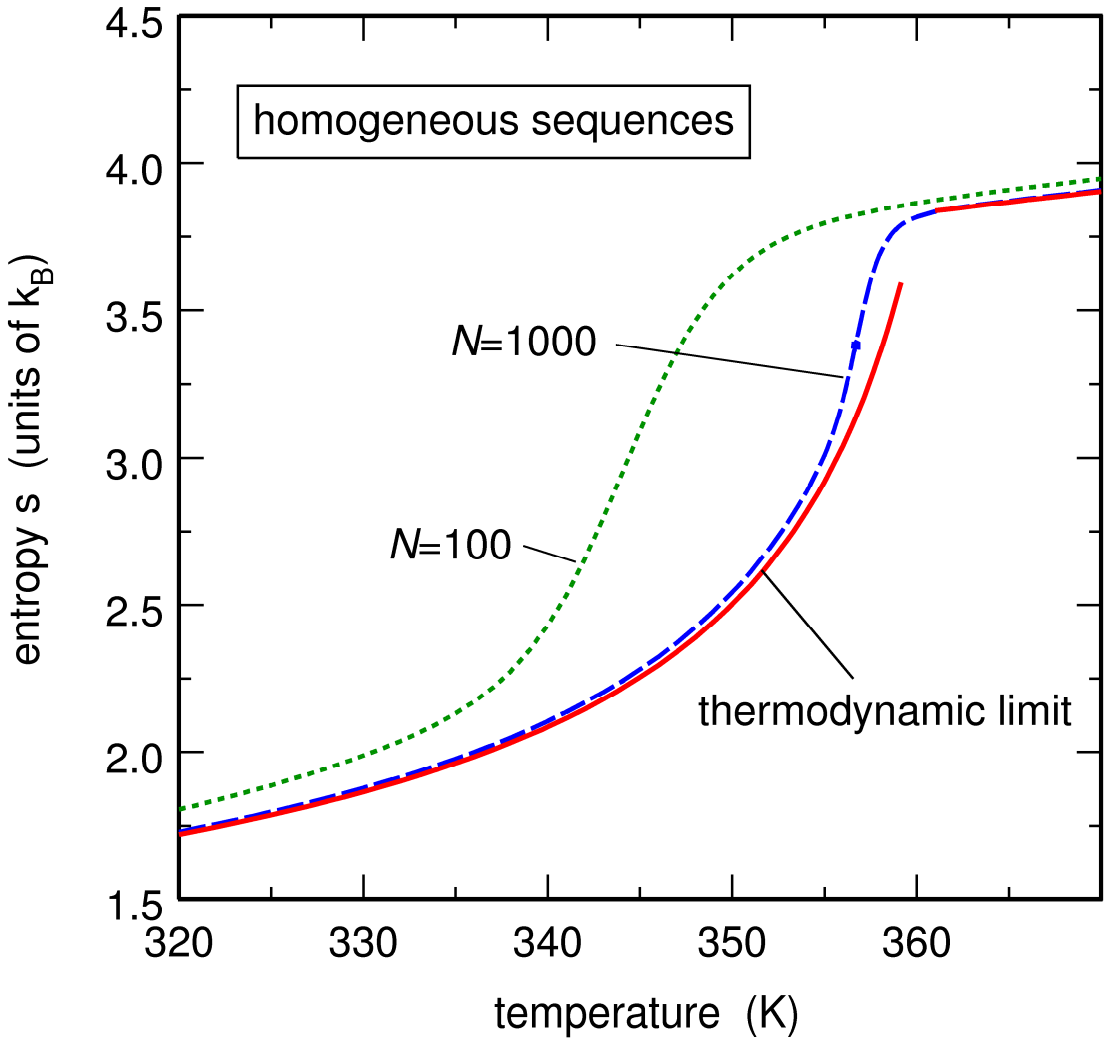

Figure 4.6*. Plot, as a function of the temperature T of the sequence, of the entropy per base pair, s, for an infinitely long homogeneous sequence and sequences with* $N = 1000$ *and* $N = 100$ *base pairs. These results were obtained from TI calculations. s is expressed in units of the Boltzmann constant* k_B*. The solid curve for the infinitely long chain is interrupted in the temperature interval where the TI method is not valid.*

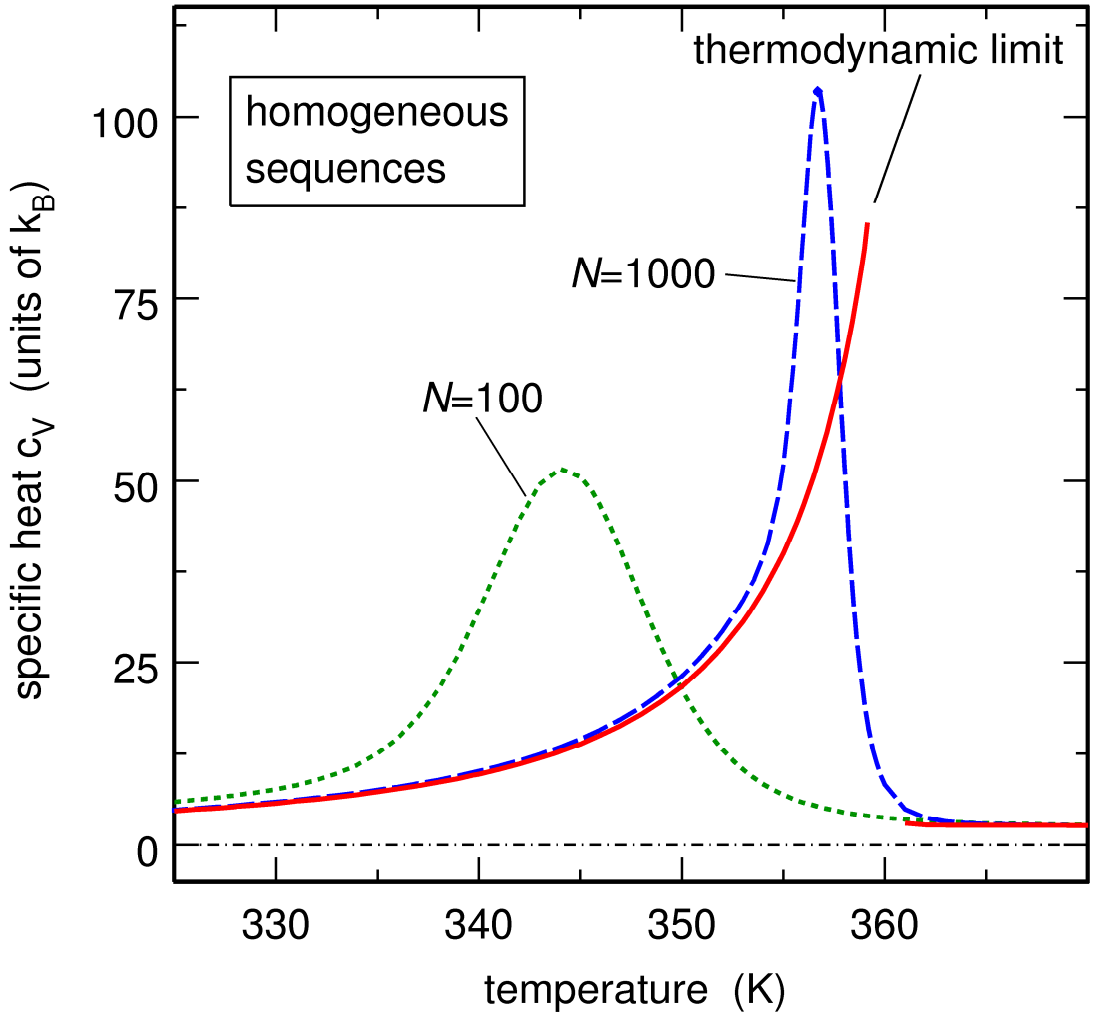

Figure 4.7. *Plot, as a function of the temperature T of the sequence, of the specific heat per base pair,* c_V, *for an infinitely long homogeneous sequence and sequences with* $N = 1000$ *and* $N = 100$ *base pairs. These results were obtained from TI calculations.* c_V *is expressed in units of the Boltzmann constant* k_B. *The solid curve for the infinitely long chain is interrupted in the temperature interval where the TI method is not valid.*

infinitely long sequences, $T_c = T_c(N = \infty)$. Here, we took advantage of the fact that the critical temperature shift $T_c - T_c(N)$ unambiguously decreases as a power of *N* and consequently found T_c as the temperature for which $\log(T_c - T_c(N))$ is best adjusted with a linear function of $\log(N)$. One gets $T_c = 359.43$ K and, as already mentioned in section 4.2, $T_c(N) \approx T_c - 539.5\,N^{-0.770}$ (see figure 4.2). Critical exponents α, β and ν are then obtained by drawing log-log plots of, respectively, c_V, $\langle y \rangle$ and ξ as a function of the temperature gap $T_c - T$ and by estimating the slope of each curve in the temperature range where the power law holds. The plots in figures 4.8 to 4.10 show that $\alpha = 1.33$, $\beta = -1.41$, and $\nu = 1.47$, not so far from the values $\alpha = 1.13$, $\beta = -1.31$, and $\nu = 1.23$ obtained with the old set of parameters [56].

The critical exponent of the specific heat, α, is thus larger than 1, which confirms that melting indeed *looks* like a first-order phase transition in the temperature range where power-laws hold. The first-order regime with $\alpha > 1$ can however not hold up to the critical temperature, because the average potential energy per base pair, $u = \langle V \rangle / N$, is expected to evolve as $u \propto (T_c - T)^{1-\alpha}$. If the regime with $\alpha > 1$ would hold up to the critical temperature, then *u* would become infinite at T_c, which is of course not possible. Figure 4.11 indeed shows that the value of α deduced from log-log plots of *u* as a function of $T_c - T$, $\alpha = 1.37$, is close to the estimation obtained from the plot of c_V, that is $\alpha = 1.33$. Most importantly, figures 4.8 to 4.11 all display a crossover from the first-order regime to another regime in the last few Kelvins below the critical temperature. We checked that the results presented in these figures are converged, that is, they do not vary when the size of the matrix in TI calculations is increased from 4201 to 8201 and the maximum value of *y* correspondingly increases from about $5000/a$ to about $10000/a$. Still, as mentioned in section 2.6, neither MD simulations nor TI calculations are able to provide a clear indication of what happens very close to T_c.

At that point, it is worth noting that analogy with the wetting transition [91] and calculations performed with a rougher model [92] suggest that the melting transition is asymptotically second-order. This actually agrees with further work performed in our group without my participation [93]. In this later work, the free energy per base pair, *f*, was separated into a singular part, f_{sing}, and a non-singular part f_{ns}. f_{ns} was taken as the free energy of two

widely separated, non-interacting DNA strands, so that f_{sing} remains constant above the critical temperature, as must be the case. Combination of equations (2.15) and (4.10) indicates that f_{sing} is expected to vary as $(T_c - T)^{2-\alpha}$ close to the critical temperature. Log-log plots of f_{sing} as a function of $T_c - T$ then led to the value $\alpha = 0.57$ for the model in equation (2.11) and the set of parameters of section 4.3 [93]. Similar calculations performed for the DPB model in equation (2.10) and increasing values of ρ confirmed that values of α estimated in this way are always comprised between 0 and 1, while values of α larger than 1 may be obtained when this critical exponent is estimated from the temperature evolution of c_V [93]. This would confirm that the melting transition is asymptotically second order (as determined from the temperature evolution of f_{sing}), while it actually looks first order (as determined from the temperature evolution of c_V) up to a few degrees below the critical temperature.

4.7. Conclusion

In this chapter, I described how we varied the parameters of the dynamical model developed in our group to get a better agreement with experimental results that were not taken into account up to now, that is, the critical force needed to keep two DNA strands separated and the sequence size dependence of the critical temperature. Then, I showed that the results obtained with the improved model agree well with those obtained from statistical models. Finally, I checked that the critical properties of the dynamical model remain qualitatively similar to those predicted with the original set of parameters. Still, it should be noted that the resolution in the temperature evolution of the specific heat per base pair, c_V, is still much too poor compared to experiment (see figure 4.5).

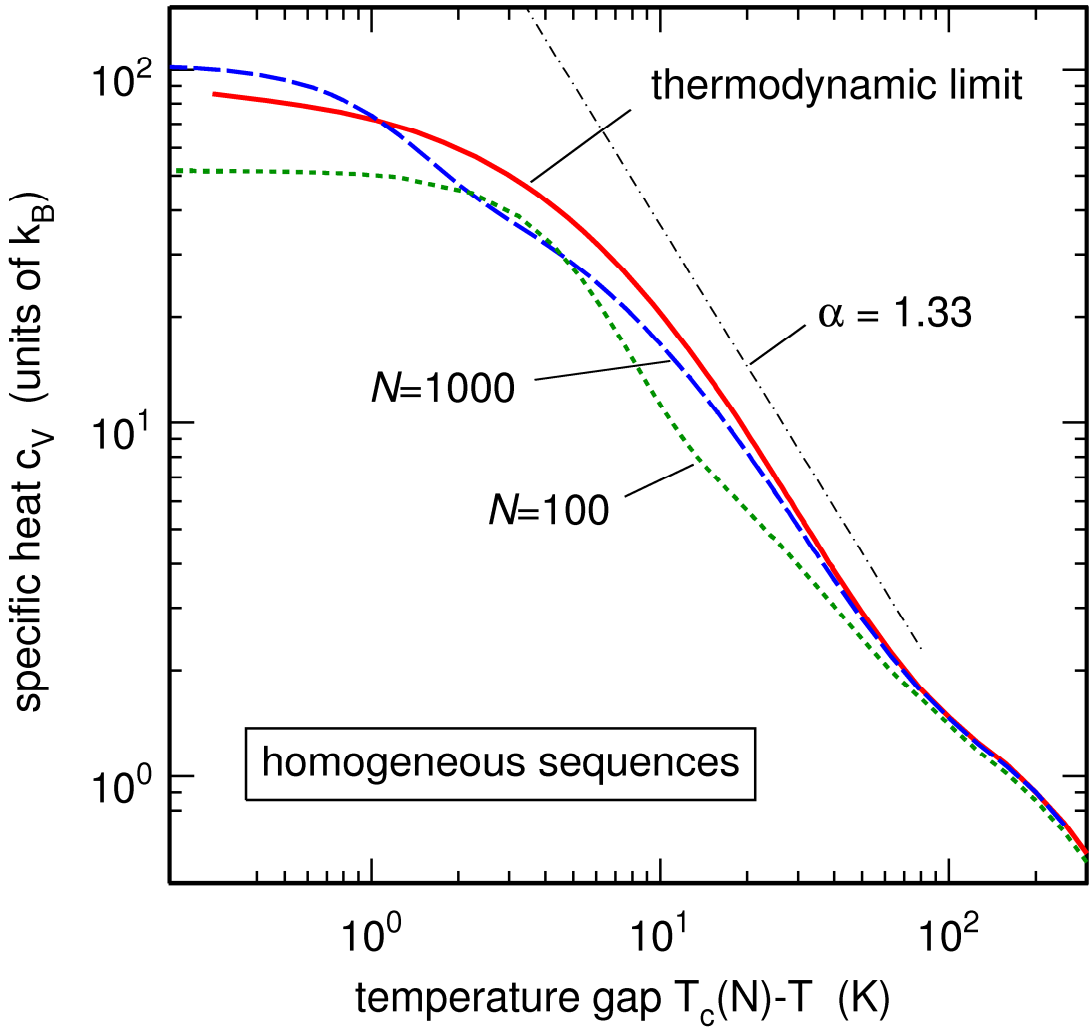

Figure 4.8. *Log-log plot, as a function of the temperature gap* $T_c(N)-T$, *of the specific heat per base pair,* c_V, *for an infinitely long homogeneous sequence and sequences with* $N=1000$ *and* $N=100$ *base pairs. These results were obtained from TI calculations.* c_V *is expressed in units of the Boltzmann constant* k_B. *The dot-dashed straight line shows the slope corresponding to a critical exponent* $\alpha=1.33$.

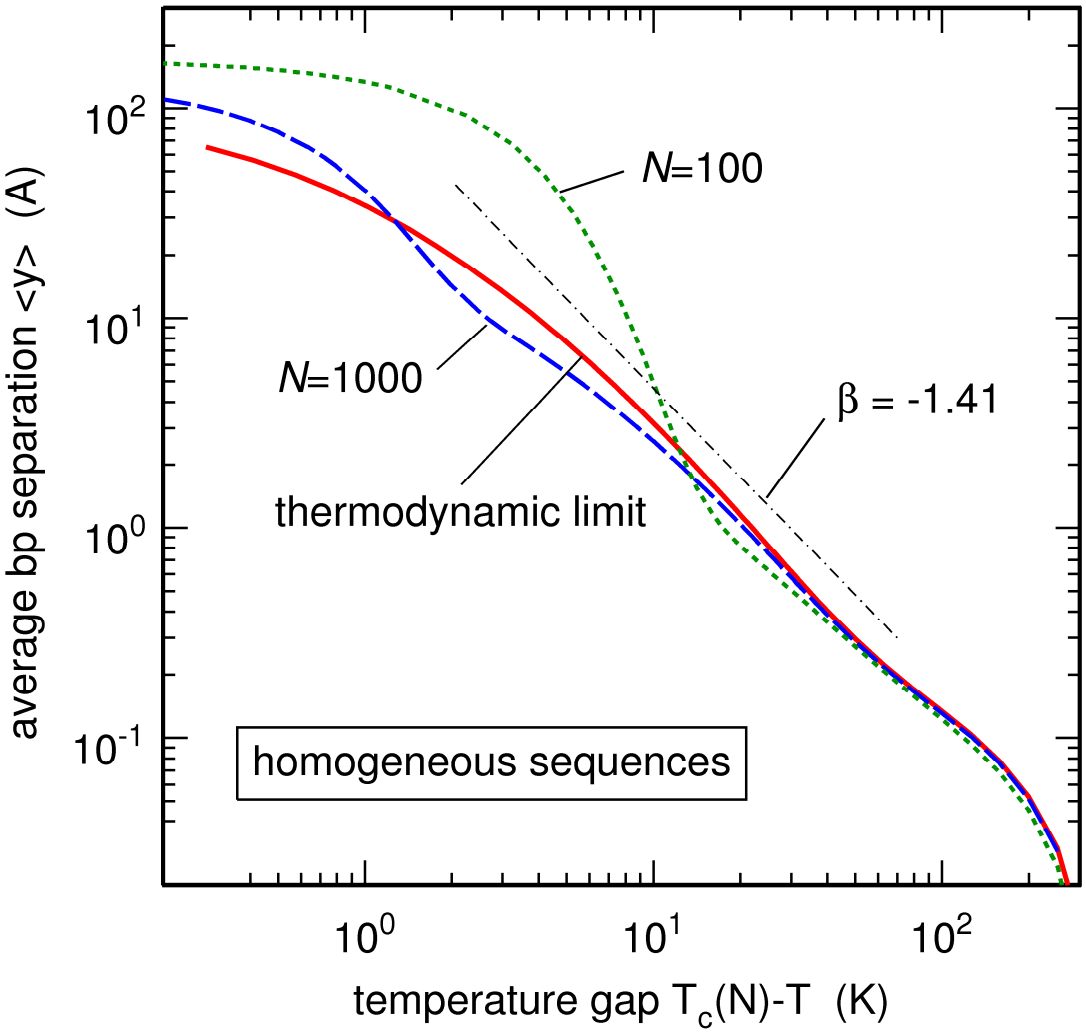

Figure 4.9. *Log-log plot, as a function of the temperature gap* $T_c(N)-T$, *of the average base pair separation,* $\langle y \rangle$, *for an infinitely long homogeneous sequence and sequences with* $N=1000$ *and* $N=100$ *base pairs. These results were obtained from TI calculations.* $\langle y \rangle$ *is expressed in Å. The dot-dashed straight line shows the slope corresponding to a critical exponent* $\beta=-1.41$.

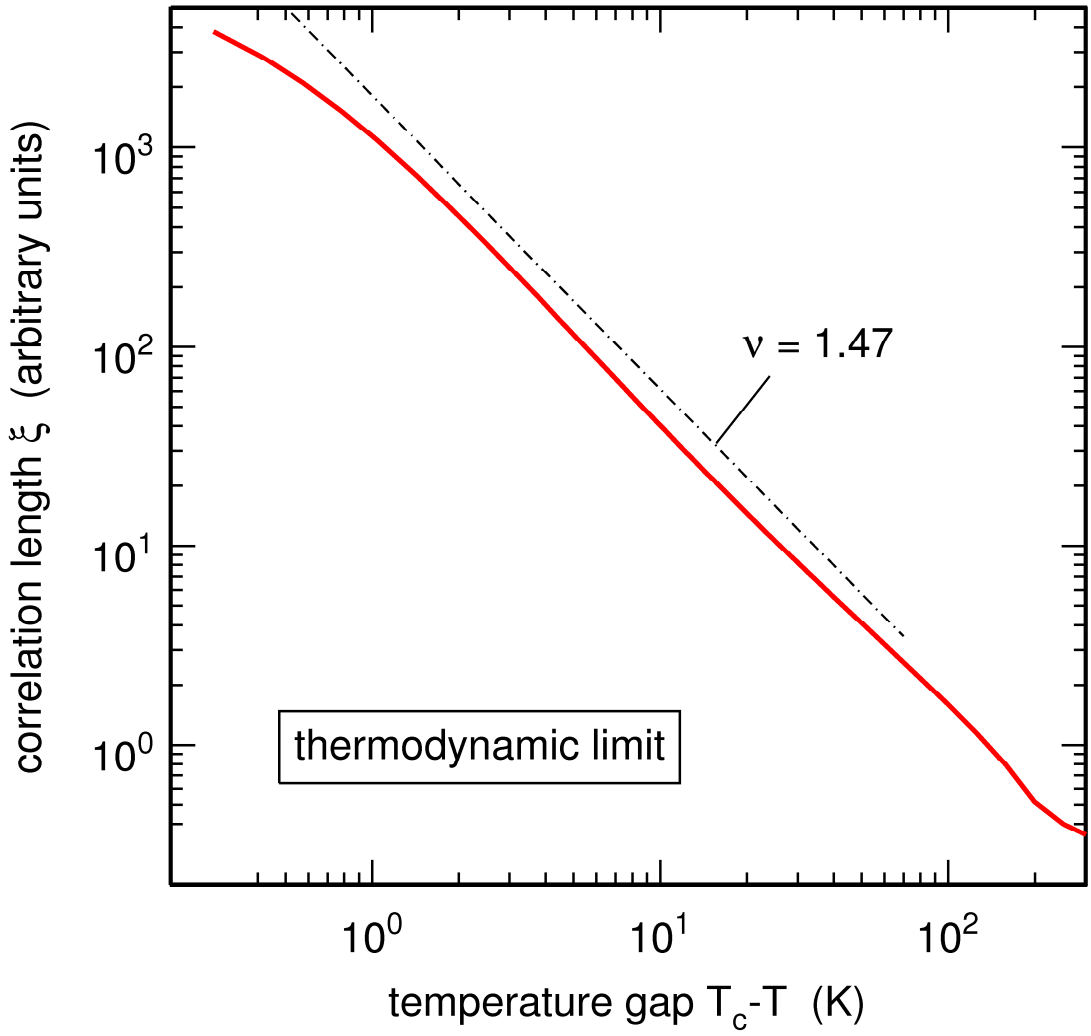

Figure 4.10. *Log-log plot, as a function of the temperature gap* $T_c - T$ *, of the correlation length,* ξ*, for an infinitely long homogeneous sequence. This result was obtained from TI calculations. The dot-dashed straight line shows the slope corresponding to a critical exponent* $\nu = 1.47$ *.*

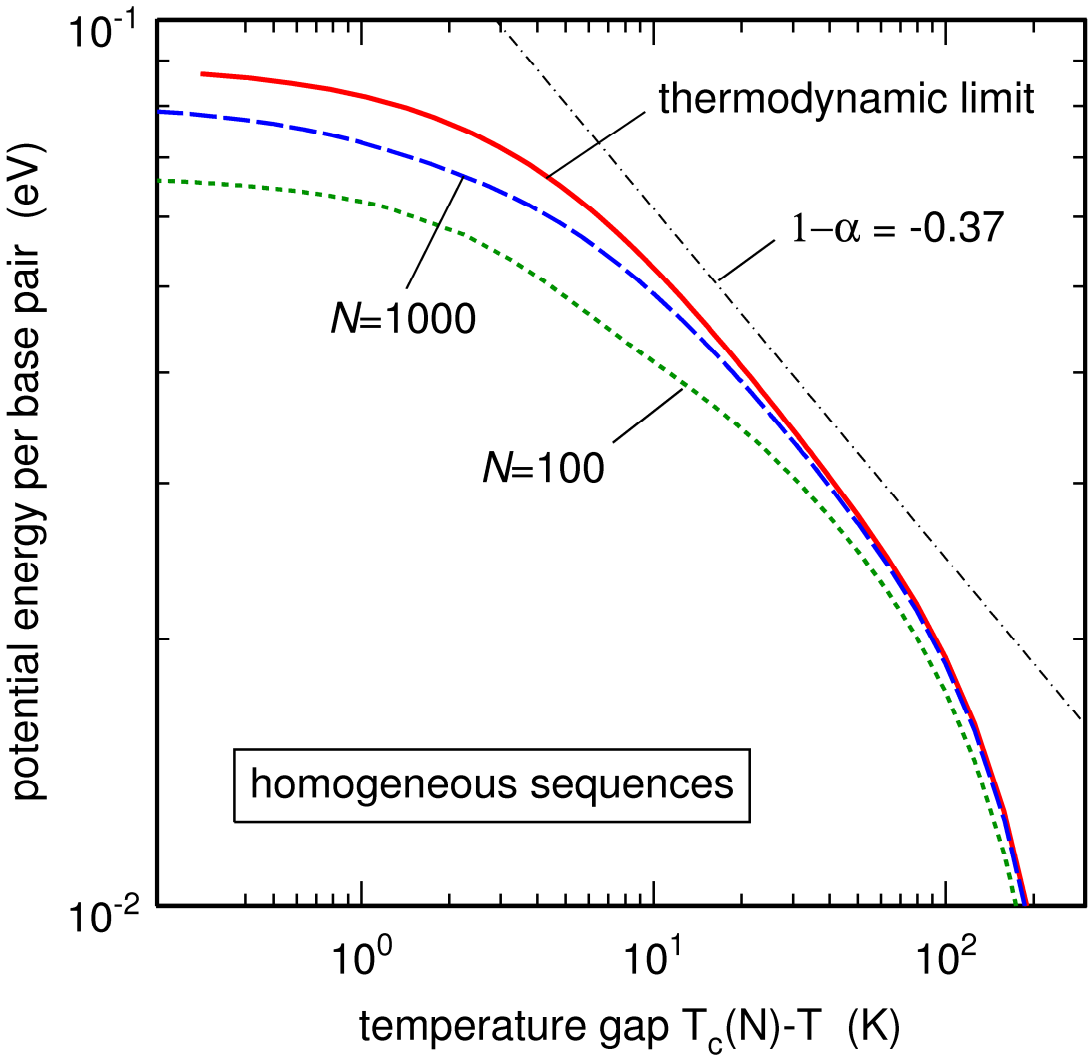

Figure 4.11. *Log-log plot, as a function of the temperature gap* $T_c(N)-T$ *, of the average potential energy per base pair,* $u=\langle V\rangle/N$ *, for an infinitely long homogeneous sequence and sequences with* $N=1000$ *and* $N=100$ *base pairs. These results were obtained from TI calculations. u is expressed in eV. The dot-dashed straight line shows the slope corresponding to a critical exponent* $1-\alpha=-0.37$ *.*

Part II: DNA – protein interactions

5. Introduction

5.1. Experiments of Riggs, Bourgeois and Cohn: the paradox of the missing salt

Although great advances have been made in genetics in the last decades and the genomes of several species are now completely mapped, there is still a lot of discussion on how gene expression takes place. Even if the steps of transcription are known, the means by which transcription factors find their targets are still not very well understood. The first and certainly one of the most important steps that have been made in this direction is due to the experiments of Riggs, Bourgeois and Cohn [10]. They measured the association rate for the *lac* repressor in various reaction conditions using a sensitive membrane filter assay for the *lac* repressor–operator complex [94]. This method consists in filtering a solution which contains repressor-operator complexes trough a membrane that only permits the passage of free DNA molecules. Therefore, they could measure how many of the DNA molecules would create complexes with the repressor in a certain amount of time. Their first experiment was performed in a buffer containing KCl, Tris-HCl and magnesium acetate at 0.01 M concentrations and they reported that the *lac* repressor binds to its operator site at a rate of about 7×10^{9} $M^{-1}s^{-1}$. This value is one to two orders of magnitude larger than what is generally assumed for diffusion limited reactions (I will explain in the next section how this theoretical rate is computed). Riggs, Bourgeois and Cohn assigned this very high rate to the existence of an electrostatic attraction between the negatively charged DNA and a positively charged site on the repressor. As an argument for this assumption they made a series of experiments where they increased the ionic concentration in the buffer up to physiological values and obtained a net decrease of the association rate down to values of the same order of magnitude as the diffusion limit (figure 5.1). The article ends with a very clear discussion of these results, from which the most important conclusion is that even though protein-DNA association reactions are accelerated by the existence of electrostatic attractions they are still diffusion limited. Another argument in support of this statement is the fact that when repeating their experiments with 20% sucrose in the buffer, the association rate is reduced by a factor of two, as expected from the change in viscosity. In the end, the authors suggest a possible mechanism for the repressor finding its target that implies “rolling” or “hopping” around the DNA sequence.

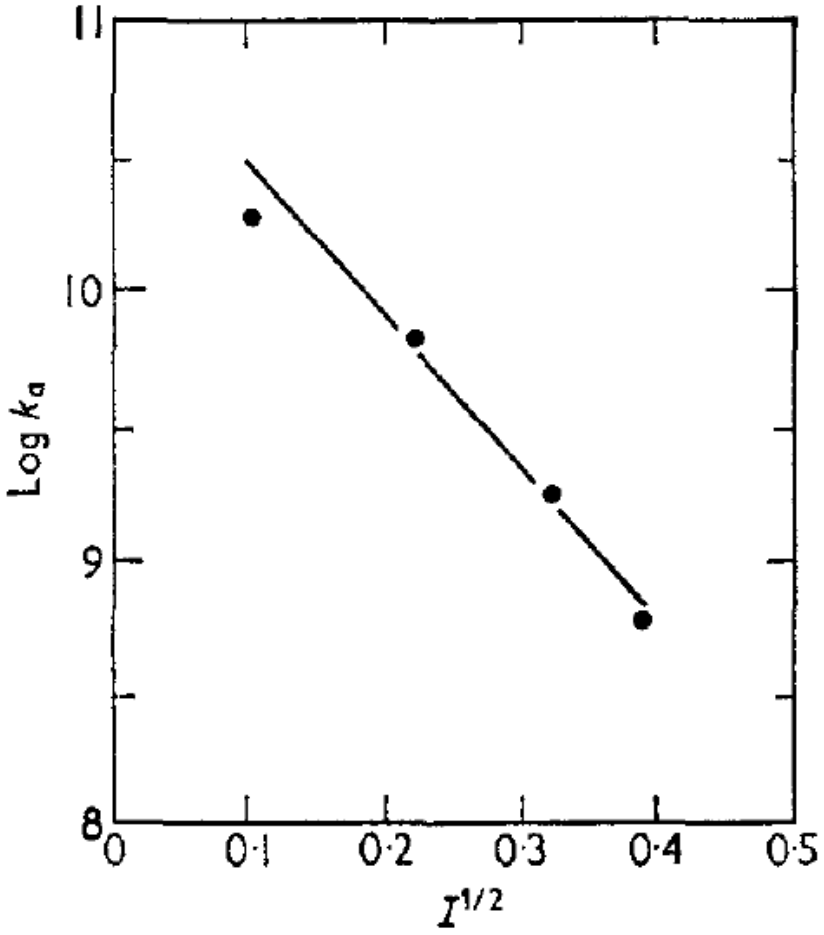

Figure 5.1. *Variation of the lac repressor-operator association rate with the KCl concentration in the buffer for the experiments of Riggs, Burgeois and Cohn in reference [10].*

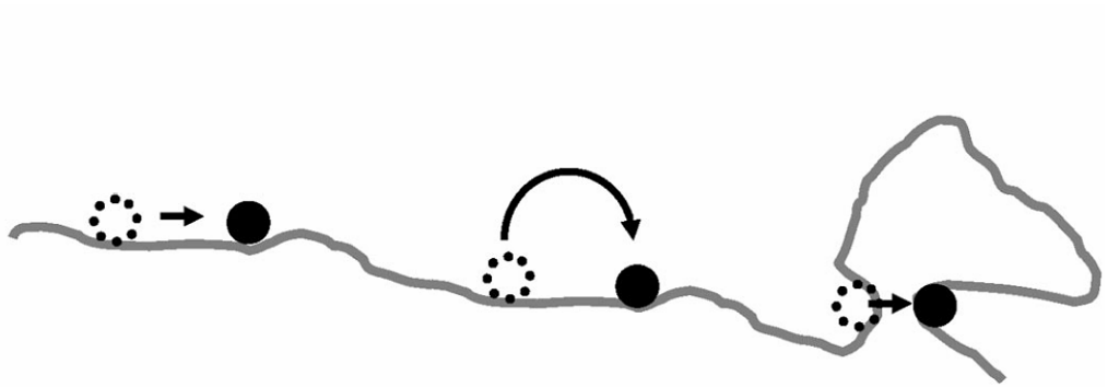

Figure 5.2. *Representation of the facilitated diffusion process. The first cartoon shows sliding, the seconds depicts hoping and the third is an image of intersegment transfer, taken from reference [16].*

This last hypothesis was later developed in a theoretical model [95] that became the inspiration for most of the models that followed. Berg, Winter and von Hippel indeed suggested that the protein's target search can be greatly accelerated by its sliding on DNA, because this would reduce the dimensionality of the process [96]. This scenario implies that the protein connects randomly on the DNA chain and then slides along it in search of its target (figure 5.2). If it does not find the target after a certain amount of time it disconnects, diffuses in the cell and then reconnects somewhere else. Besides this, some proteins can also do intersegment transfer, which implies transiently doubly binding to two DNA segments. The combination of these processes is now known as facilitated diffusion, and is the basis for many theoretical models that aim to describe protein-DNA interactions [15-19]. Most of them are analytical models inspired by the theoretical description of chemical reactions, which only take into consideration the kinetics of the entire molecule population of the system. These models usually imply assuming the values of non-specific association and dissociation probabilities in order to compute the specific association rate.

Surprisingly enough, the majority of the theoretical works tend to assign the very high association rates measured in the early experiment of Riggs *et al* to facilitated diffusion and they seem to ignore the fact that this rate decreases when increasing the ionic strength of the buffer. This negligence has propagated, giving birth to a general conception that facilitated diffusion accelerates DNA-protein association with a factor as high as 100 compared to the diffusion limit.

The development of new techniques for *in vivo* microscopy has recently permitted direct visualization of the motion of proteins inside the cell, the precise determination of their diffusion coefficients and has also evidenced the existence of facilitated diffusion [98-109]. These experiments show that the values of the one-dimensional diffusion coefficient of the protein are up to a thousand times smaller than those for three-dimensional diffusion, and that the protein spends more time sliding than diffusing through the buffer. These facts should cast a doubt on the belief that facilitated diffusion is necessarily faster than normal diffusion. Incidentally, a recent review of experimental results by Halford[21] states that there is "no known example of a protein binding to a specific DNA site at a rate above the diffusion limit", and that "the rapidity of these reactions is due primarily to electrostatic interaction between oppositely charged molecules". This work clearly reminds us that the very high association rate observed in the first experience of Riggs, Bourgeois and Cohn is due to the absence of salt in the buffer, and that once these

experiments are repeated at a higher salt concentration the association rate greatly decreases. It also points out that these results have been confirmed by several other experiments [22-24]. Also, Halford suggests that we should put "an end to forty years of mistakes in DNA-protein association kinetics". Therefore, these experimental results indicate that the debate around how proteins find their targets should now concentrate on whether facilitated diffusion really accelerates DNA sampling and which are the conditions for this to happen.

One of the goals of my thesis is to propose a mechanical model for the study of non-specific DNA protein interactions and to use it to discuss in what conditions the alternation between one-dimensional diffusion and three-dimensional diffusion can lead to faster DNA sampling than normal diffusion. It implies developing a Hamiltonian that describes the interactions inside the cell and solving the equations of motion for the particles in the system. The next chapters contain a general presentation of existing models, a description of the model proposed, a study of what characteristics of the protein affect the speed of DNA sampling and a comparison of the results presented here with existing theoretical results.

5.2. The diffusion limit and the Smoluchowski rate

The purpose of this section is to explain how the diffusion limit is computed in the case of proteins binding on DNA.

In a reaction where one spherical molecule A associates with a spherical molecule B, the association rate constant will reach the diffusion limit when every collision of A with B will result in a complex. The rate constant for a diffusion limited reaction is usually computed using the Smoluchowski equation [110]:

$$k_{\mathrm{Smol}} = 4\pi(1000\, N_{\mathrm{Avog}})(D_{\mathrm{A}} + D_{\mathrm{B}})(r_{\mathrm{A}} + r_{\mathrm{B}}) \tag{5.1}$$

It is expressed in units of $M^{-1}s^{-1}$. N_{Avog} is Avogadro's number, D_{A} and D_{B} are the 3D diffusion constants of the colliding species A and B (in units of $m^2\ s^{-1}$), and r_{A} and r_{B} their reaction radii (in meters). This is deduced by computing the particle flux that diffuses trough a spherical molecule in the case of the steady state solutions for Fick's equation [111]. In order to show the steps of this procedure, I will give here the case of a reaction with a fixed target, but the

deduction is similar for all other cases. The concentration of particles diffusing in space at the distance *r* from an absorbing particle of radius *a* is given by:

$$C(r) = C_0(1 - \frac{a}{r}) \tag{5.2}$$

where C_0 is the initial concentration of particles. The flux trough the spherical absorber is:

$$J(r) = -D\frac{\partial C}{\partial r} = -DC_0\frac{a}{r^2} \tag{5.3}$$

The particles are absorbed by the sphere with a rate *I* that is equal to the area times the inward flux, which equals -*J(a):*

$$I = 4\pi DaC_0 \tag{5.4}$$

The reaction rate is given by the coefficient that relates *I* and C_0. If the molecules are not considered spherical, then the problem becomes more complicated because the Smoluchowski rate also depends on a geometrical factor, which is proportional to the fraction of the number of collisions in which the two molecules face each other in the correct orientation for complex formation. When applying equation (5.1) to the particular case of a protein associating with DNA, Riggs *et al* [10] considered that the reaction radius $r_A + r_B$ is of the order of 0.5 nm, which is approximately the size of a base or an amino-acid. They further on pointed out that the diffusion constant of DNA is negligible compared to that of the protein and consequently estimated that $D_A + D_B \approx 0.50\times10^{-10}$ m^2 s^{-1}, on the basis of the 150000 molecular weight of the *lac* repressor (this is very close to the value obtained from Einstein's formula for the diffusion constant of a sphere). By plugging these numerical values in equation (5.1), one obtains $k_{Smol} \approx 2\times10^8$ M^{-1} s^{-1}, which is about 35 times less than the value measured in reference [10]. However, the crucial point is that equation (5.1) is valid only if molecules A and B have no net charge or if these charges are neutralized by counterions [112]. If this is not the case, then the association rate for free diffusion must be modified to include an electrostatic factor f_{elec} [21,95]:

$$k = k_{Smol} f_{elec} \tag{5.5}$$

f_{elec} is larger than 1 if the interacting surfaces of A and B possess opposite charges. It is instead comprised between 0 and 1 if the sign of the charges is the same. Moreover, f_{elec} usually tends towards 1 when the ionic strength is increased, because there are more and more counterions that neutralize the charges on the interacting surfaces of A and B. If the salinity of the buffer is very

low, it can be considered in a first approximation that the electrostatic interaction between DNA and the protein is an unscreened Coulomb potential $q_A q_B/(4\pi\varepsilon r)$, where q_A and q_B are the charges on interacting particles A and B, and r is the distance between them. Debye [113] showed that, in this case, f_{elec} is given by:

$$f_{\text{elec}} = \frac{x}{e^x - 1} \tag{5.6}$$

and

$$x = \frac{q_A q_B}{4\pi\varepsilon (r_A + r_B) k_B T} \tag{5.7}$$

By plugging $q_A = -5\bar{e}$ (the DNA electrostatic charge for about 7 bps), $q_B = 10\bar{e}$ (the typical value for a protein effective charge [114,115]), $r_A + r_B = 0.5$ nm, and $\varepsilon = 80\varepsilon_0$ in equations (5.6) and (5.7), one obtains $f_{\text{elec}} \approx 70$, which is of the same order of magnitude as the decrease in the association rate constant that Riggs *et al* measured when increasing the salinity of the buffer up to physiological values (0.1 M KCl instead of 0.01 M) [10]. As far as I know, there does not exist such an explicit formula as equation (5.6) neither for the screened Debye-Hückel potential, nor for the sum of a screened Debye-Hückel potential and an excluded volume term, as the one I use in my model (see below). It was, however, checked numerically that the association rate for a screened Debye-Hückel potential is comprised between Schmolukowski's rate and Debye's one [116].

5.3. Kinetic models

Almost all the models for the description of protein-DNA association that were developed until now are based on similar series of assumptions regarding the nature of site targeting. These models assume that proteins alternate between three-dimensional (3D) motion in the cell and one-dimensional (1D) sliding along the DNA chain and that both motions are purely diffusive with known diffusion coefficients D_{1D} and D_{3D}. These values are used to compute various observables such as the binding rate and the total time required for targeting as a function of a set of well defined geometric quantities, such as the DNA sequence length L and the average volume V of the cell, and a more or less extended list of rate constants and reaction probabilities (see, for

example, table I of reference [95]). For example, Halford and Marko [16] computed the reaction rate starting from the probability for a protein to find a nearby target by diffusion. They divided this process into several stages: first, the protein has to diffuse in solution until it encounters a DNA coil. Once the protein enters the coil it has a certain probability to find the target. Therefore, it has to visit the coil several times in order to be sure that it connects to the specific site. Defining the sliding length as the starting distance for which the probability of binding is 0.5, and then applying the laws for 1D and 3D diffusion to the steps of the targeting process, Halford and Marko showed that the reaction rate for unit protein concentration can be expressed as:

$$k = \left(\frac{1}{D_{3D}\,\ell_{sl}} + \frac{L\,\ell_{sl}}{D_{1D}\,V} \right)^{-1} \tag{5.8}$$

Since they considered the 3D diffusion-limited rate to be equal to $a\,D_{3D}$, where a is the size of the target, it follows that the acceleration of the reaction due to facilitated diffusion is $k/(a\,D_{3D})$. After a couple of additional hypotheses regarding the values of these parameters (of which the most important is that $D_{1D} \approx D_{3D}$), Halford and Marko concluded that this ratio is at its maximum equal to about 30 for an optimal sliding length $\ell_{sl} \approx 100$ base pairs, a value that is close to those obtained from single-molecule experiments [102,104,109]. This model gives a good hint on the qualitative dependence of the targeting speed on many parameters, but it is not very accurate. For example, it neglects the numerical factors that are present in the 1D and 3D diffusion equation and consequently also the 4π factor in the association rate.

Most of the other kinetic models rely heavily on terms or expressions that are quite difficult to relate to experimentally measured properties of molecules and/or quantities derived from dynamical systems (for example various correlation terms that are supposed to arise when the protein switches from 1D to 3D motion). Therefore, I have only found one kinetic model to which it is possible and meaningful to compare the results obtained with the dynamical model I have developped. Using older calculations of Szabo *et al* [117], Klenin, Merlitz, Langowski and Wu derived from first principles that the mean time of the first arrival of the protein at the target of radius a can be written in the form [118]:

$$\tau = \left(\frac{V}{8D_{3D}\xi} + \frac{\pi L\xi}{4D_{1D}}\right)\left[1 - \frac{2}{\pi}\arctan\left(\frac{a}{\xi}\right)\right] \tag{5.9}$$

where D_{1D} and D_{3D} are the diffusion coefficients of the protein in the buffer and along the DNA sequence, respectively, and ξ

$$\xi = \sqrt{\frac{1}{2\pi}\frac{V}{L}\frac{D_{1D}}{D_{3D}}\frac{\tau_{1D}}{\tau_{3D}}} \tag{5.10}$$

the distance where the efficiencies of the two types of diffusion become equal to each other (τ_{1D} and τ_{3D} are the average times the protein spends in the bound and free states, respectively). The accuracy of equation (5.9) was checked for the simple system where the protein is described as a random walker that is allowed to enter freely in the neighborhood of the DNA but has a given finite probability to exit this volume at each time step [18,118]. Although equation (5.9) was developed starting from a different idea than that leading to equation (5.8), it gives the same general tendencies for the reaction time (it has the same qualitative dependence on the diffusion coefficients for example). As will be shown below, all the quantities that appear in equations (5.9) and (5.10) can also be derived using molecular dynamics, and the mean time of first arrival τ can be related to the rate constant k. Further on, equation (5.9) will therefore be used for all comparisons between kinetic models and the model described in this work.

5.4. The volume of the Wiener sausage

Before describing the dynamical model proposed in this work, I would like to make a short synthesis of some mathematical results regarding Brownian motion and random walks. A pure Brownian process is diffusive, so it is characterised by a diffusion coefficient D such that:

$$< \mathbf{R}^2 > (t) = 2dDt \tag{5.11}$$

where d is the dimension of the space. The spatial region travelled by a spherical Brownian particle in a time t is formed by all points within a fixed distance of the centre of the particle. In mathematical terms, this comes to compute the Lebesgue measure of the space covered by the Brownian motion. This is also known as the Wiener sausage [119] (this name is a pun resulting from a combination between “Wiener processes”, which are a class of mathematical processes that include Brownian motions and which were named after the mathematician who studied them, and “Viennese sausages”, which can be used as a spatial representation of the volume covered by a 3D random walker). The volume of the Wiener sausage is important for the analysis of most of

the physical processes that can be described by a random walk. The analytical expressions for its long time asymptotic values in the case of a diffusing sphere of radius δ are [120,121]:

$$\ell(t) \approx \sqrt{\frac{16}{\pi} D_{1D}\ t} \tag{5.12}$$

in the case of a 1D random walk,

$$S(t) \approx \frac{2t\pi}{\log(t)} \tag{5.13}$$

for the surface covered in a 2D process and

$$V(t) \approx 4\pi\ \delta\ D_{3D}\ t \tag{5.14}$$

for the volume spanned by a 3D Brownian motion.

Equation (5.12) shows, as one would intuitively expect from the diffusion equation, that the length covered by a 1D random walker increases as the square root of time. However, the result for the 3D case is less intuitive. Even though the average distance travelled by a 3D diffusing particle increases as a square root of time, the volume it covers increases linearly. The essential reason for this is that in three dimensions a random walk has a zero probability to re-visit a point in space, in contrast to 1D and 2D motions. It is therefore not surprising that the visited volume increases more rapidly in three dimensions than in the 1D or 2D cases.

6. The dynamical model: description and first results

In this chapter, I describe the model I have developed to investigate DNA-protein interactions and the first basic results obtained therewith.

6.1. Description

I consider a system composed of a cell (or its nucleus) described as a sphere of radius R_0 containing a protein and several DNA segments. For the description of DNA, I have chosen an existing wormlike chain model [20] inspired from polymer physics. This is a bead-and-spring model that accurately reproduces the DNA molecule's persistence length and translational diffusion coefficient. The DNA segments consist of chains of *n* beads connected by springs, which stand for fifteen base pairs. Each bead has a hydrodynamic radius $a_{\mathrm{DNA}} = 1.78\,\mathrm{nm}$ and an electrostatic charge of $e_{\mathrm{DNA}} = 0.243\times10^{10} l_0 \bar{e} \approx 12\bar{e}$ ($\bar{e}$ is the charge of the electron) placed in its center. The equilibrium value for the inter-bead distance is $l_0 = 5\,\mathrm{nm}$. I chose the number of beads in a segment, *n*, in such a way that the length of each DNA segment is approximately equal to the radius of the cell, thus filling the cell homogeneously with DNA but avoiding a chain's excessive curvature. This description does not take into account histones or other proteins that may be connected to DNA, but it has been shown that the sliding track of bacterial DNA can be truncated into short and mostly uniformly distributed DNA segments [122]. The number *m* of segments is chosen so that the density of bases inside the cell is close to real values. As pointed out in [16], the volume *V* of the cell and the total DNA length *L* are connected according to $V = w^2 L$, where *w* represents roughly the spacing of nearby DNA segments. *m* must therefore fulfil the relation $\frac{4}{3}\pi R_0^3 \approx w^2 m n l_0$, where the average value $w = 45.0$ nm holds for both prokaryote and eukaryote cells [16]. However, this model best describes organisms where DNA is not packed in chromatin, like, for example, viruses. In this first work, I used three different sizes for the system (in order to make sure that the results are not size dependent):

- $m = 30$ segments of $n = 33$ beads (i.e. a total of 14850 base pairs) and a cell radius $R_0 = 0.134$ μm

- $m = 50$ segments of $n = 40$ beads (i.e. a total of 30000 base pairs) and a cell radius $R_0 = 0.169$ μm
- $m = 80$ segments of $n = 50$ beads (i.e. a total of 60000 base pairs), with the cell radius $R_0 = 0.213$ μm

and $w = 45$ nm in the three cases. Figure 6.1 is an image of the cell for the last case.

The potential energy E_{pot} of the system consists of three terms:

$$E_{\text{pot}} = V_{\text{DNA}} + V_{\text{DNA/prot}} + V_{\text{wall}} \tag{6.1}$$

where V_{DNA} describes the potential energy of the DNA beads and the interactions between them, $V_{\text{DNA/prot}}$ stands for the interactions between the protein bead and DNA segments, and V_{wall} models the interactions with the cell wall, which refrain all the beads in the system from going outside the cell. V_{DNA} is taken from reference [20]:

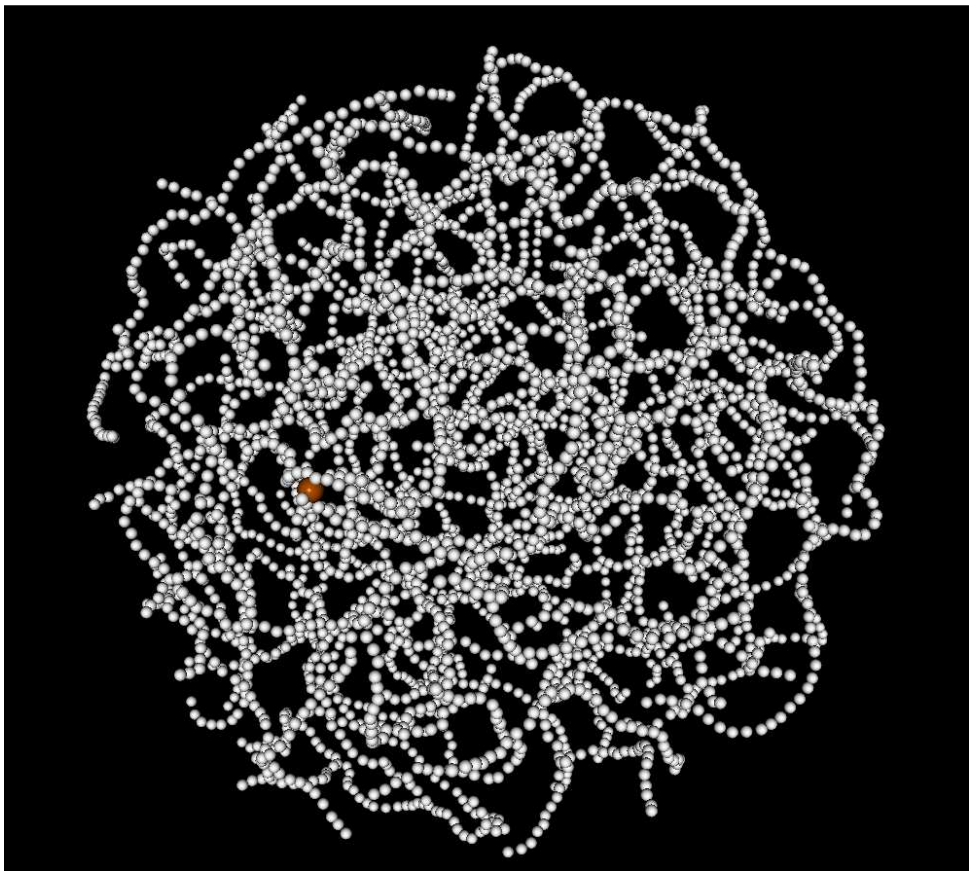

Figure 6.1. *Snapshot of the cell for the case where* $m = 80$ *and* $n = 50$. *In white are the DNA beads and in dark orange is the protein.*

$$V_{\mathrm{DNA}} = E_{\mathrm{s}} + E_{\mathrm{b}} + E_{\mathrm{e}}$$

$$E_{\mathrm{s}} = \frac{h}{2}\sum_{j=1}^{m}\sum_{k=1}^{n-1}\left(l_{j,k} - l_0\right)^2$$

$$E_{\mathrm{b}} = \frac{g}{2}\sum_{j=1}^{m}\sum_{k=1}^{n-2}\theta_{j,k}^2$$

$$E_{\mathrm{e}} = \frac{e_{\mathrm{DNA}}^2}{4\pi\varepsilon}\sum_{j=1}^{m}\sum_{k=1}^{n-2}\sum_{K=k+2}^{n}\frac{\exp\left(-\frac{1}{r_D}\left\|\mathbf{r}_{j,k}-\mathbf{r}_{j,K}\right\|\right)}{\left\|\mathbf{r}_{j,k}-\mathbf{r}_{j,K}\right\|}$$

$$+\frac{e_{\mathrm{DNA}}^2}{4\pi\varepsilon}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{J=j+1}^{m}\sum_{K=1}^{n}\frac{\exp\left(-\frac{1}{r_D}\left\|\mathbf{r}_{j,k}-\mathbf{r}_{J,K}\right\|\right)}{\left\|\mathbf{r}_{j,k}-\mathbf{r}_{J,K}\right\|}\quad , \tag{6.2}$$

where $\mathbf{r}_{j,k}$ denotes the position of bead k of segment j, $l_{j,k} = \left\|\mathbf{r}_{j,k} - \mathbf{r}_{j,k+1}\right\|$ the distance between two successive beads belonging to the same segment, and $\theta_{j,k}$ the angle formed by three successive beads on the same segment

$$\cos\theta_{j,k} = \frac{\left(\mathbf{r}_{j,k} - \mathbf{r}_{j,k+1}\right)\cdot\left(\mathbf{r}_{j,k+1} - \mathbf{r}_{j,k+2}\right)}{\left\|\mathbf{r}_{j,k} - \mathbf{r}_{j,k+1}\right\|\left\|\mathbf{r}_{j,k+1} - \mathbf{r}_{j,k+2}\right\|} \tag{6.3}$$

E_{s} is the bond stretching energy. This is actually a computational device without real biological meaning, which is essentially aimed at avoiding dealing with rigid rods. The stretching force constant is fixed at $h = 100\, k_{\mathrm{B}}T / l_0^2$, with $T = 298$ K. This value was chosen in order to balance between using a time step that is as large as possible and having only small displacements from the equilibrium length. For this value of h one gets $< l > / l_0 = 1.02$.

E_{b} is the elastic bending potential. There are several methods to approximate the bending rigidity constant g of a worm-like chain, all aiming to give a correct persistence length. One of the simplest is:

$$g = \frac{pk_BT}{l_0}, \tag{6.4}$$

where p is the persistence length (here $p = 50.0$ nm, i. e. 10 beads) and l_0 is the inter-particle distance. This would give a value of $g = 10\, k_{\mathrm{B}}T$, but here I have used the value $g = 9.82\, k_{\mathrm{B}}T$, which I borrowed from references [20,123].

E_e is a Debye-Hückel potential, which describes repulsive electrostatic interactions between DNA beads [20,124,125]. This potential also takes into considerations the screening of interactions due to the ions in the buffer, so that in equation (6.2) $r_D = 3.07$ nm stands for the Debye length at 0.01 M molar salt concentration of monovalent ions:

$$r_D = \sqrt{\frac{\varepsilon k_B T}{2 N_A e^2 I}}, \tag{6.5}$$

where $\varepsilon = 80\,\varepsilon_0$ is the dielectric constant of the buffer, N_A is Avogadro's number, and I is the ionic strength of the buffer:

$$I = \frac{1}{2}\sum_{i=1}^{n} c_i z_i^2 , \tag{6.6}$$

where n is the number of types of ions in the buffer, c_i is their concentration and z_i is their charge. Electrostatic interactions between neighbouring beads belonging to the same segment are not included in the expression of E_e in equation (6.2), because it is considered that these nearest-neighbour interactions rather contribute to the stretching and bending terms.

The potential V_{wall}, which models the interactions between DNA and the protein and the cell wall, is taken as a sum of short range repulsive terms that act on the beads that trespass the radius of the cell, R_0, and repel them back inside the cell:

$$V_{wall} = k_B T \sum_{j=1}^{m}\sum_{k=1}^{n} f\left(\left\|\mathbf{r}_{j,k}\right\|\right) + 10\, k_B T\; f\left(\left\|\mathbf{r}_{prot}\right\|\right) \tag{6.7}$$

where $\mathbf{r}_{prot}$ denotes the position of the protein and f is a function defined as:

if $x \le R_0$: $f(x) = 0$

if $x > R_0$: $f(x) = \left(\frac{x}{R_0}\right)^6 - 1$. (6.8)

The coefficients $k_B T$ and $10\,k_B T$ in equation (6.7) were roughly adjusted by hand, in order that, at 298 K and for cell radii R_0 comprised between 0.134 and 0.213 μm, all the beads (DNA and protein) remain confined inside a sphere of radius $\approx 1.10\,R_0$, which insures that the time spent by the beads outside the cell is negligible. The coefficient is 10 times larger for the protein than for DNA, because the protein is modelled by a single bead, so that its mobility is much larger than

that of the interconnected DNA beads and its motion outside the sphere of radius R_0 more difficult to oppose.

The interaction $V_{\text{DNA/prot}}$ between the protein and DNA beads is the sum of an attractive and a repulsive term:

$$V_{\text{DNA/prot}} = E_{\text{e}}^{(\text{P})} + E_{\text{ev}}$$

$$E_{\text{e}}^{(\text{P})} = -\frac{e_{\text{DNA}} e_{\text{prot}}}{4\pi\varepsilon} \sum_{j=1}^{m} \sum_{k=1}^{n} \frac{\exp\left(-\frac{1}{r_D} \left\| \mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}} \right\| \right)}{\left\| \mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}} \right\|} \tag{6.9}$$

$$E_{\text{ev}} = k_{\text{B}} T \frac{e_{\text{prot}}}{e_{\text{DNA}}} \sum_{j=1}^{m} \sum_{k=1}^{n} F\left(\left\| \mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}} \right\| \right)$$

where F is a function defined as:

$$\text{if } x \le \sqrt{2}\,\sigma : F(x) = 4\left(\left(\frac{\sigma}{x} \right)^4 - \left(\frac{\sigma}{x} \right)^2 \right) + 1$$

$$\text{if } x > \sqrt{2}\,\sigma : F(x) = 0 \tag{6.10}$$

and $\sigma = a_{\text{DNA}} + a_{\text{prot}} = 5.28$ nm. $E_{\text{e}}^{(\text{P})}$ is the Debye-Hückel potential, which models the attractive electrostatic interactions between the protein and DNA beads, while E_{ev} is an excluded volume term, which prevents the protein bead from superposing to a DNA bead and $E_{\text{e}}^{(\text{P})}$ from diverging. E_{ev} is sometimes taken as the repulsive part of the Lennard-Jones potential [126]. Being of order 12, this function is however so sharp that it leads too often to numerical bugs, while the order 4 function $F(x)$ enables relatively trouble-free calculations. The prefactor of E_{ev} was chosen as $k_{\text{B}} T\, e_{\text{prot}} / e_{\text{DNA}}$, because this insures that the DNA/protein interaction $V_{\text{DNA/prot}}$ displays a global minimum very close to $\sigma = a_{\text{DNA}} + a_{\text{prot}}$, whatever the charge e_{prot} of the protein bead (figure 6.2). Intuitively, $V_{\text{DNA/prot}}$ must indeed be minimum at some value close to the sum of the radii of DNA and the protein (that is, close to σ) in order for sliding to take place. Moreover, I will take advantage of the fact that the position of this minimum does not depend on e_{prot} to let e_{prot} assume different values, thereby varying the percentage of time the protein bead spends in 1D sliding and 3D motion.

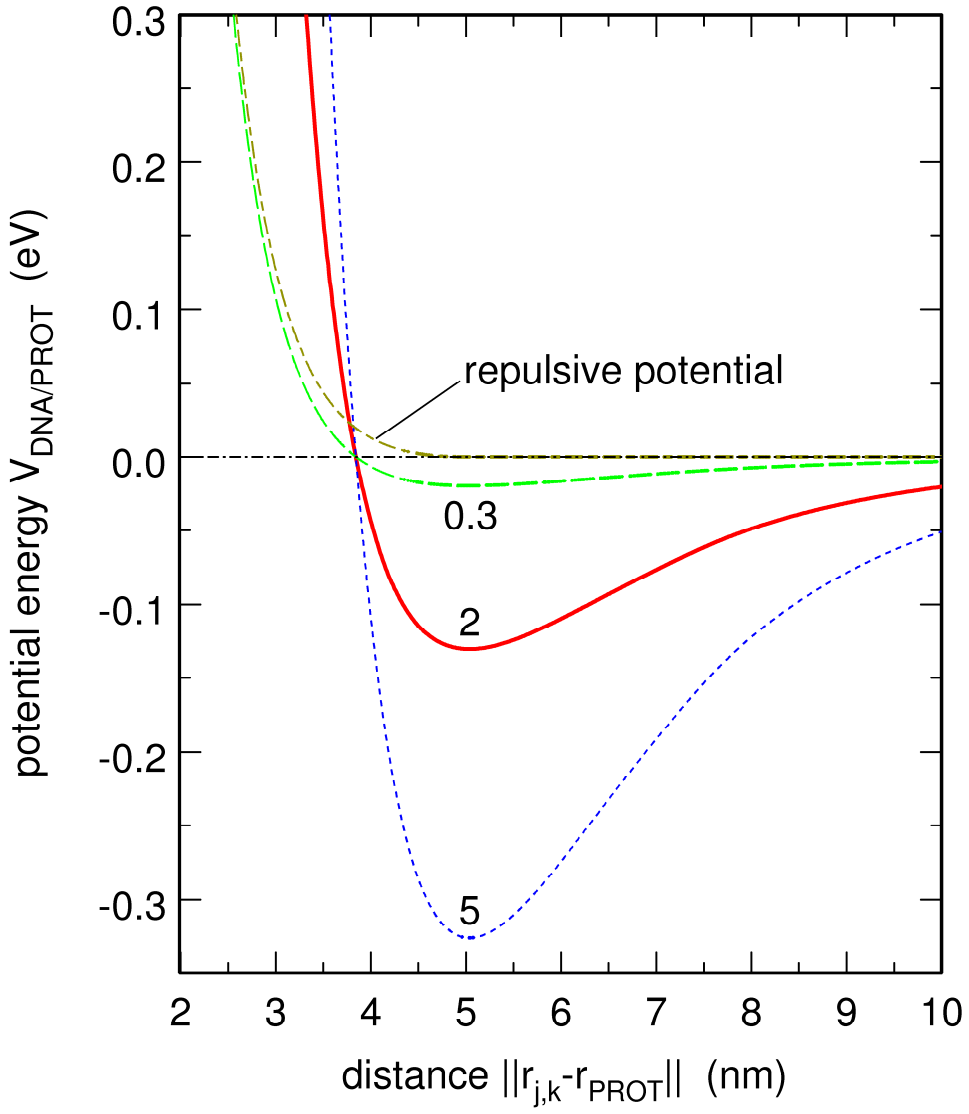

Figure 6.2. *Plot, as a function of the distance* $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{\text{prot}}\right\|$ *between the centres of the two beads, of the interaction potential* $V_{\text{DNA/prot}}$ *between the protein bead and bead k of DNA segment j, for three different values of* $e_{\text{prot}}/e_{\text{DNA}}$ *(0.3, 2 and 5) and a purely repulsive potential, which is just the repulsive part of the potential with* $e_{\text{prot}}/e_{\text{DNA}}=0.3$. $V_{\text{DNA/prot}}$ *is expressed in eV and* $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{\text{prot}}\right\|$ *in nm. Note that the three curves with* $e_{\text{prot}}/e_{\text{DNA}}=0.3$, *2 and 5 all display a minimum located at* $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{\text{prot}}\right\|$ *=5.04 nm, close to* $\sigma=a_{\text{DNA}}+a_{\text{PROT}}=5.28$ *nm.*

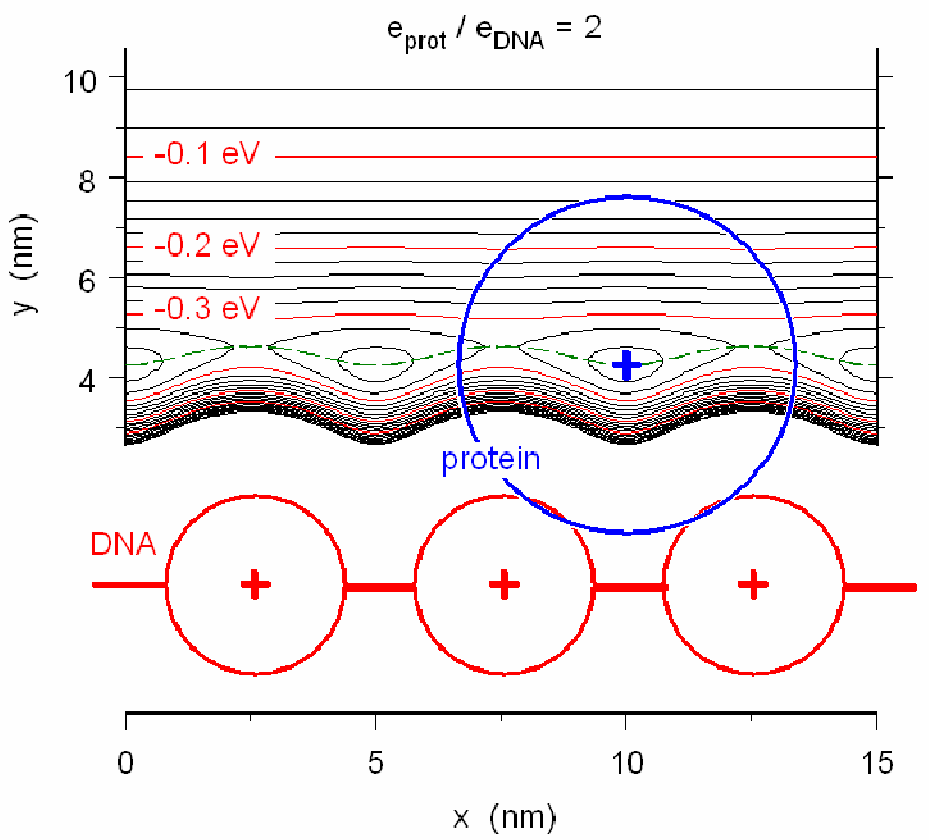

Figure 6.3. *Plot of the equal energy lines for the DNA-protein interaction potential with* $e_{\text{prot}} / e_{\text{DNA}} = 2$. *At infinite DNA-protein separation the potential energy is 0. Contour lines are separated by 0.02 eV, with the lines in red corresponding to -0.1, -0.2 and -0.3 eV. The green line denotes the minimum energy path. The hydrodynamic radii of DNA and the protein beads are indicated, with the protein sitting on a minimum.*

Figure 6.3 shows the lines of equal potential for the interaction energy between the protein and a fixed, straight DNA chain for $e_{\text{prot}} / e_{\text{DNA}} = 2$. Except for the depth of the minimum, equipotential lines for values of $e_{\text{prot}} / e_{\text{DNA}}$ ranging from 0.3 to 5 are very similar to figure 6.3. It is important to notice that the potential well is deeper by a factor up to almost 3 if the protein interacts simultaneously with several DNA beads. Also, the potential barrier the protein has to pass in order to slide from one bead to the other is very small compared to the maximum depth of the potential wall, so it does not have any significant effect on the sliding motion for moderate values of e_{prot}. In contrast, it cannot be excluded that this barrier plays a significant role in the subdiffusive behaviour, which is observed for larger values of e_{prot} or highly deformable proteins (see chapters 7 and 8).

For describing the motion of the system I used Brownian dynamics. Brownian dynamics is employed in molecular dynamics simulations when one wants to avoid dealing with explicit solvent molecules. Instead, molecular collisions are accounted for by adding random forces to the potential and friction forces in a Newtonian motion. The Brownian dynamics equation of motion for an ensemble of particles is:

$$\mathbf{M}\ddot{\mathbf{r}} = -\nabla E(\mathbf{r}(t)) - \mathbf{Z}\dot{\mathbf{r}}(t) + \xi(t) \tag{6.11}$$

where **M** is a diagonal matrix containing the masses of the particles, *E* is the potential energy, **Z** is the tensor containing friction coefficients, and *ξ(t)* are random forces with mean and covariance given by:

$$\begin{aligned} &<\xi(t)>= 0 \\ &<\xi(t)\xi(t')^T>= 2k_B T\mathbf{Z}\delta(t-t') \end{aligned} \tag{6.12}$$

Equation (6.12) is based on the fluctuation-dissipation theorem. This theorem states that, for a randomly moving particle, friction is related to a random force. Since the random force does not have a time scale, the time scale of the motion of such a system is given by the inertial relaxation times, defined as the inverses of the eigenvalues of the matrix $\mathbf{M}^{-1}\mathbf{Z}$. When these times are short compared to the timescale of the simulation it is possible to ignore inertia and assume that $\mathbf{M}\ddot{\mathbf{r}}(t) = 0$. Then, equation (6.11) becomes:

$$\dot{\mathbf{r}}(t) = -\mathbf{Z}^{-1}\nabla E(\mathbf{r}(t)) + \mathbf{Z}^{-1}\xi(t), \tag{6.13}$$

which can also be written as:

$$\dot{\mathbf{r}}(t) = -\frac{\mathbf{D}}{k_B T}\nabla E(\mathbf{r}(t)) + \xi_B(t) \tag{6.14}$$

where **D** is the diffusion tensor, connected to the friction coefficients by:

$$\mathbf{D} = k_B T\mathbf{Z}^{-1} \tag{6.15}$$

The mean and covariance of the random forces are connected to the diffusion tensor through:

$$\begin{aligned} &\xi_B(t) = 0 \\ &<\xi_B(t)\xi_B(t')^T>= 2\mathbf{D}\delta(t-t') \end{aligned} \tag{6.16}$$

The algorithm that I have chosen for solving these equations is that of Ermack and McCammon [127]. This algorithm is based on the approximation that momentum relaxation occurs much faster than position relaxation. This condition translates in a condition that the time step is sufficiently large: $\Delta t > M / 6\pi\eta a$ (for a detailed explanation see reference [127]). In this case, this

gives a lower limit for the time step of 1 ps. According to the first-order version of this algorithm, the updated position vector for the beads is given by:

$$\mathbf{r}^{(n+1)} = \mathbf{r}^{(n)} + \frac{\Delta t}{k_{\mathrm{B}}T}\mathbf{D}^{(n)}.\mathbf{F}^{(n)} + \sqrt{2\Delta t}\ \mathbf{L}^{(n)}.\xi_G^{(n)} \tag{6.17}$$

where Δt is the time step. Note that $\mathbf{r}^{(n)}$ and $\mathbf{r}^{(n+1)}$ are collective vectors that include the position vectors $\mathbf{r}_{j,k}$ of all DNA beads, as well as the position vector $\mathbf{r}_{\mathrm{prot}}$ of the protein bead, at steps *n* and *n*+1. The second term in the right-hand side of equation (6.17) models the diffusive effects of the buffer. $\mathbf{F}^{(n)}$ is the collective vector of inter-particle forces arising from the potential energy E_{pot} and $\mathbf{D}^{(n)}$ the hydrodynamic interaction diffusion tensor. As in [126], I built the successive tensors $\mathbf{D}^{(n)}$ using a modified form of the Rotne-Prager tensor for unequal size beads [128-130] (see equations (26)-(28) of [126]). The third term in the right-hand side of equation (6.17) models the effects on $\mathbf{r}^{(n+1)}$ of collisions between the buffer and the protein and DNA beads. $\xi_G^{(n)}$ is a vector of random numbers extracted at each step *n* from a Gaussian distribution of mean 0 and variance 1, while $\mathbf{L}^{(n)}$ is the lower triangular matrix obtained from the Choleski factorization of $\mathbf{D}^{(n)}$:

$$\mathbf{D}^{(n)} = \mathbf{L}^{(n)}.{}^{t}\mathbf{L}^{(n)} \tag{6.18}$$

where ${}^{t}\mathbf{L}^{(n)}$ denotes the transpose of $\mathbf{L}^{(n)}$. The CPU time required to factor the diffusion matrix increases as the cube of the number of beads that are taken into account in $\mathbf{D}^{(n)}$, so that the Choleski factorization of $\mathbf{D}^{(n)}$ turns out to be the limiting step for the investigation of the dynamics of large systems. There is an algorithm that can be used to decrease the exponent from 3 to 2.25 [131,132], but I chose to use a more drastic approximation. Since the main purpose of this work is to study the interaction between DNA and the protein, it is most important that the motion of DNA close to the protein is modelled correctly. The results are little affected if the motion of DNA far from the protein is handled in a cruder way. Therefore, I used equations (6.17) and (6.18) to calculate the position at each time step of the protein and the 100 DNA beads closest to it, while the positions of the remaining DNA beads were obtained from the diagonal approximation of equation (6.17), that is

$$\mathbf{r}^{(n+1)} = \mathbf{r}^{(n)} + \frac{\Delta t}{6\pi\eta\ a_{\mathrm{DNA}}}\mathbf{F}^{(n)} + \sqrt{\frac{2\ k_B T\ \Delta t}{6\pi\eta\ a_{\mathrm{DNA}}}}\ \xi_G^{(n)} \tag{6.19}$$

where $\eta = 0.00089$ Pa s denotes the viscosity of the buffer at 298 K. Note that equation (6.19) is just the first-order discretization of the usual Langevin equation without hydrodynamic interactions and with the second-order term arising from kinetic energy dropped. When considering a system with 2000 DNA beads, use of equations (6.17) and (6.18) to update the positions of the protein and the 100 closest DNA beads slows down calculations by only 10% compared to the case where equation (6.19) is used for all beads. In contrast, the CPU time is already multiplied by a factor larger than 2 if equations (6.17) and (6.18) are used for the 200 DNA beads closest to the protein. On the other hand, I checked that use of equation (6.19) to update the position of all beads leads to results that differ substantially from those presented further on, while use of equations (6.17) and (6.18) to update the position of the 200 DNA beads closest to the protein, instead of the 100 closest ones, leads to similar results. The use of equation (6.17) for the protein and the 100 closest DNA beads and of equation (6.19) for the other DNA beads therefore appears as a very reasonable choice.

For all simulations, the *m* DNA segments were first placed inside the cell according to a randomization procedure that insures an essentially uniform distribution of the beads in the cell (see figure 6.4). The protein bead was then placed at random in a sphere of radius $R_0 / 5$. In order to avoid too strong repelling interactions at time $t = 0$, all initial configurations where the distance between the protein and at least one DNA bead turned out to be smaller than $\sigma = a_{\mathrm{DNA}} + a_{\mathrm{prot}} = 5.28$ nm were rejected. The equations of motion (6.17) and (6.19) were then integrated for 10 μs, in order for the system to equilibrate at the correct temperature. The quantities of interest were subsequently obtained by integrating the equations of motion for longer time intervals and averaging over several different trajectories. Finally, I have checked that time steps Δt equal to 25, 100 and 400 ps lead to identical results (see figure 6.5). Most of the results discussed in this chapter were consequently obtained with $\Delta t = 100$ ps, although a few ones dealing with the system with 4000 DNA beads were obtained with $\Delta t = 400$ ps.

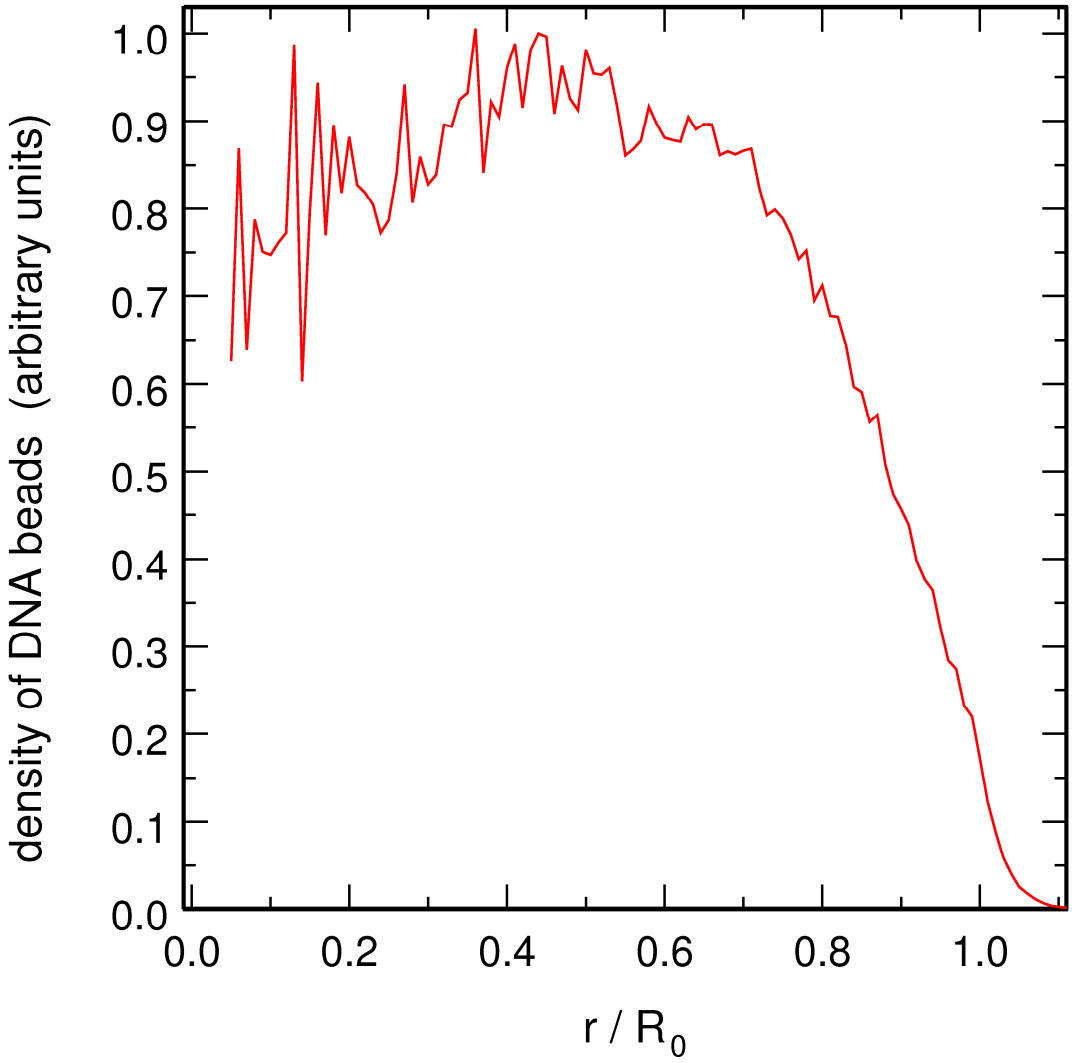

Figure 6.4. *Profile of the number of DNA beads per unit volume as a function of the distance r from the centre of the cell after an integration time of 30 µs. The maximum of the curve was arbitrarily scaled to 1. This profile was averaged over 64 different trajectories with 2000 DNA beads.*

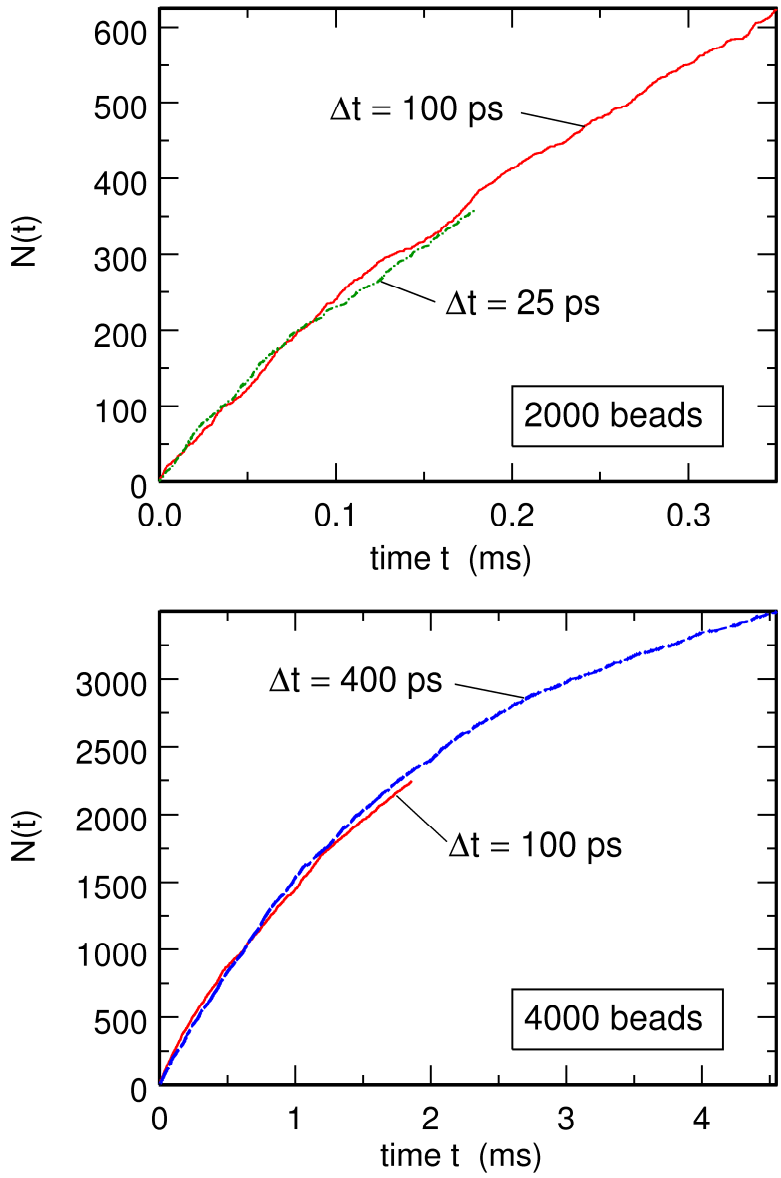

Figure 6.5*. Comparison of results obtained with different time steps. Both plots show the evolution of* $N(t)$*, the number of different DNA beads visited by the protein at time t. It is considered that a DNA bead and the protein are in contact if the distance between the centres of the two beads is smaller than* $\sigma = a_{\mathrm{DNA}} + a_{\mathrm{prot}} = 5.28$ *nm. The top plot shows the evolution of* $N(t)$ *for the system with 2000 DNA beads,* $e_{\mathrm{prot}} / e_{\mathrm{DNA}} = 1$ *and time steps* $\Delta t = 25$ *and 100 ps. The bottom plot shows the evolution of* $N(t)$ *for the system with 4000 DNA beads,* $e_{\mathrm{prot}} / e_{\mathrm{DNA}} = 1$ *and time steps* $\Delta t = 100$ *and 400 ps. Each curve was averaged over 6 different trajectories.*

6.2. 1D and 3D diffusion and DNA sampling

For studying the nature of the motion of the protein and of the DNA sampling process, I chose three quantities: the proportion ρ_{1D} of the total simulation time (excluding equilibration) that the protein spends connected to the DNA chain, which is equivalent to the probability of sliding, the number $N(t)$ of different DNA beads it has visited at time *t*, and the number n_{sim} of DNA beads to which it is simultaneously connected when it is not diffusing freely in the buffer. Actually, the number $N(t)$ of sites visited by a diffusing particle is the most important parameter for studying a search process: it is indeed connected to the association rate in diffusion controlled reactions and to the first passage time on a given site. On the other hand, examination of n_{sim} is useful when varying the charge of the protein, because it may indicate when the values of the parameters become unrealistic.

For the repulsive $V_{DNA/prot}$ potential displayed in figure 6.2, the DNA and the protein never attract each other. The protein therefore moves almost freely in the buffer, except that it is repelled by the excluded volume interaction E_{ev} whenever the distance to a DNA bead becomes too small. This case is actually very close to a diffusion limited reaction. Because of the large density of DNA beads, the probability for the protein to be found close to a DNA bead is not negligible: if one considers that the protein interacts with bead *k* of DNA segment *j* when $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\right\| \le \sigma$, then DNA "fills" about 3% of the cell volume and the protein is expected to spend approximately the same amount of time interacting with DNA, in spite of the absence of attractive interactions. This is indeed the case, as can be checked in figure 6.6, which shows the proportion of time ρ_{1D} during which the protein interacts with a DNA bead as a function of the ratio e_{prot}/e_{DNA}. In this plot, the points at $e_{prot}/e_{DNA}=0$ correspond to the repulsive potential in figure 6.2, while circles and lozenges respectively denote results obtained with the $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\right\| \le \sigma$ and $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\right\| \le 1.5\,\sigma$ criteria for interacting beads. It is seen that ρ_{1D} is indeed close to 3% for the repulsive potential and the $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\right\| \le \sigma$ criterion.

I found that the number of beads visited by the protein in the absence of the attractive part of $V_{\mathrm{DNA/prot}}$ increases with time following an exponential law:

$$\frac{N(t)}{mn} = 1 - \exp\left(-\kappa \frac{t}{mn}\right) \tag{6.20}$$

where $\kappa = 1.09\ \mu s^{-1}$ (see figure 6.7). The most important aspect of this law is that it implies that $N(t)$ increases linearly with a rate κ as long as N remains sufficiently small compared to the total number mn of DNA beads inside the cell, while this rate steadily decreases down to zero when N comes closer and closer to mn, due to saturation (there are less and less new beads to visit).

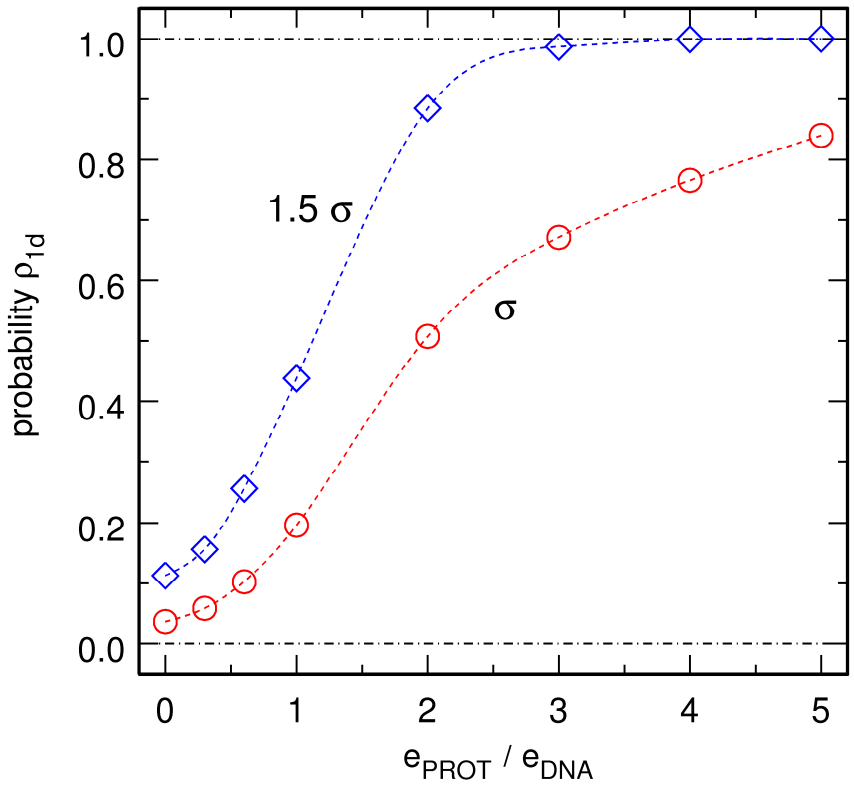

Figure 6.6. *Plot, as a function of the ratio $e_{\mathrm{prot}}/e_{\mathrm{DNA}}$, of the portion of time $\rho_{1\mathrm{D}}$ during which the protein remains attached to a DNA bead. The abscissa axis actually corresponds to the variation of e_{prot} at constant e_{DNA}. Circles and lozenges denote results obtained with, respectively, the $\|\mathbf{r}_{j,k} - \mathbf{r}_{\mathrm{prot}}\| \le \sigma$ and $\|\mathbf{r}_{j,k} - \mathbf{r}_{\mathrm{prot}}\| \le 1.5\,\sigma$ criterions for interacting beads. The point at $e_{\mathrm{prot}}/e_{\mathrm{DNA}} = 0$ was obtained with the repulsive potential of figure 6.2. Each point was averaged over 12 different trajectories propagated for 100 μs for the system with 2000 beads.*

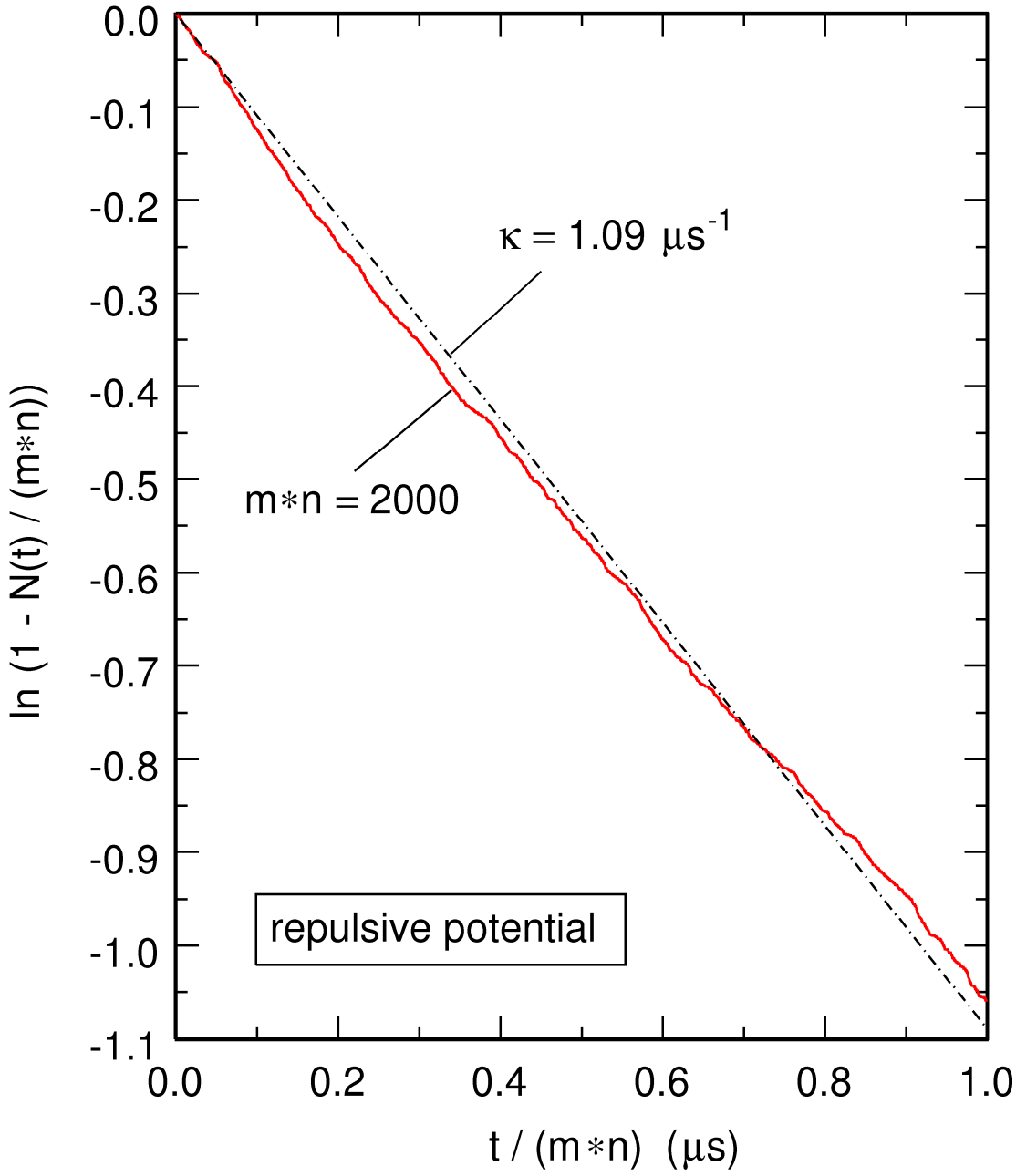

Figure 6.7. *Plot of* $\ln(1-N(t)/(mn))$ *as a function of* $t/(mn)$ *for the system with* $mn=2000$ *DNA beads and the repulsive potential of figure 6.2. The dot-dashed straight line represents the same plot for the expression of* $N(t)$ *in equation (6.20) and a rate* $\kappa=1.09\ \mu s^{-1}$.

To compare this law with equation (5.14) one firstly needs to remember that DNA is homogenously distributed in the cell and that its motion is slow compared to that of the protein. Then *N(t)* and *V(t)* can be related through $N(t) = cV(t)$. In the end, we can deduce from equation (5.14) an increase rate for the number of visited sites:

$$\kappa = 4\pi D_{3D} \delta c \tag{6.21}$$

Knowing that is sufficient for the protein to touch a DNA site to consider that is has been visited, one can approximate that the volume covered by the protein in a time *t* is equivalent to the trace of a 3D random walk performed by a sphere with radius $\delta = a_{\text{prot}} + a_{\text{DNA}}$. One can therefore compute the association rate from equation (5.14) according to:

$$\kappa \approx 4\pi \left(a_{\text{prot}} + a_{\text{DNA}}\right) D_{3D}\ c \tag{6.22}$$

When plugging in equation (6.22) the actual concentration of DNA beads, $c = 9.89 \times 10^{22}$ beads/m^3, and the 3D diffusion coefficient at 298 K of a sphere of radius a_{prot},

$$D_{3D} = \frac{k_B T}{6\pi \eta\, a_{\text{prot}}} \approx 0.7 \times 10^{-10}\ \text{m}^2/\text{s} \tag{6.23}$$

one obtains $\kappa \approx 0.46$ beads/μs, which is less than a factor 2 away from the value of κ obtained from the simulations, and coincides almost perfectly with the value that is obtained when the positions of all the beads are updated according to equation (6.19), that is when hydrodynamic interactions are completely disregarded. I will come back to this point in chapter 8.

If $e_{\text{prot}} / e_{\text{DNA}} > 0$, then the interaction potential $V_{\text{DNA/prot}}$ between the protein and DNA beads displays a minimum close to $\sigma = a_{\text{DNA}} + a_{\text{prot}}$ (figure 6.2), so that the motion of the protein results from the balance of conflicting constraints: $V_{\text{DNA/prot}}$ tends to localize the protein close to DNA segments, while stochastic interactions with the buffer tend to release the protein bead in the bulk of the cell. Figure 6.6 indicates that the motion of the protein therefore consists of a combination of 1D sliding and 3D motion for values of $e_{\text{prot}} / e_{\text{DNA}}$ not too large, up to $e_{\text{prot}} / e_{\text{DNA}} \approx 3$. For larger values of $e_{\text{prot}} / e_{\text{DNA}}$, the electrostatic attraction between the protein and DNA is predominant, so that the protein spends most of the time in the neighbourhood of a DNA segment. The ratio $e_{\text{prot}} / e_{\text{DNA}} \approx 1$ corresponds to an effective protein charge $e_{\text{prot}} \approx 12\, \bar{e}$, which

is of the same order of magnitude as experimentally determined protein effective charges [114,115].

At this point, it should be mentioned that hydrodynamic interactions tend to decrease the ratio of time the protein spends sliding along the DNA chain compared to 3D motion in the buffer. For example, if one neglects all hydrodynamic interactions, then ρ_{1D} is found to be equal to 0.60 (respectively, 0.95) for $e_{prot}/e_{DNA}=1$ and the $\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\|\le\sigma$ criterion (respectively, the $\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\|\le 1.5\sigma$ criterion) for interacting beads, instead of $\rho_{1D}=0.20$ and 0.44 when hydrodynamic interactions are taken into account. As will be discussed in chapter 8, this has marked consequences on the number $N(t)$ of different beads visited by the protein at time *t*.

Figure 6.8 illustrates the typical trajectory of a protein bead for the ratio $e_{prot}/e_{DNA}=1$. During the 15 µs time interval displayed in this figure, the protein visits four different segments. Globally, sliding along each segment can last several µs, but it is frequently interrupted by shorter time intervals during which the protein is released in the buffer and at the end of which it reattaches to the same segment either at the same position or at a neighbouring one. These short jumps are often called "hops" [95,109,133]. On the other hand, the protein sometimes moves almost freely and for longer time intervals (several µs) in the buffer before reattaching to another segment or eventually to the same segment but at a rather different position. Note also that intersegmental transfer, which involves an intermediate state where the protein is simultaneously bound to two different segments [95,109,133], is observed in these simulations, especially at larger values of e_{prot}/e_{DNA}, although this kind of motion is not illustrated in figure 6.8.

It can easily be checked that the number of different DNA beads visited by the protein during 1D diffusion very precisely follows the square root law in equation (5.12). For example, the solid line in figure 6.9 shows the evolution of $N(t)$ for the system with 2000 DNA beads and $e_{prot}/e_{DNA}=1$, obtained by averaging over 43 sliding events, which lasted more than 1 µs and during which the protein neither detached from the DNA segment for more than 0.07 µs nor reached one of the extremities of the segment. It can be seen that this solid curve very closely

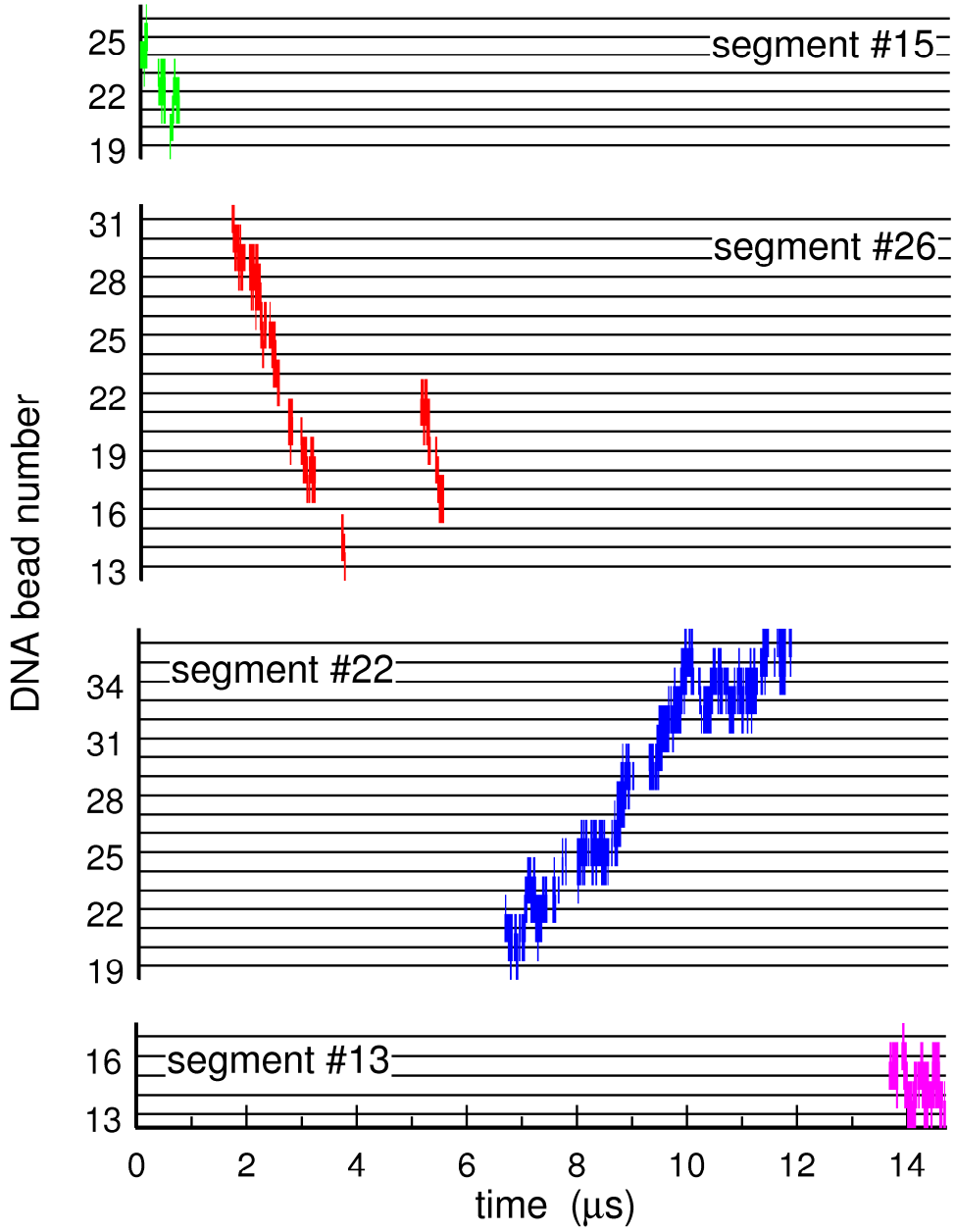

Figure 6.8. *Typical protein trajectory for the system with 2000 DNA beads and the ratio* $e_{\text{prot}} / e_{\text{DNA}} = 1$. *This plot indicates, at each time, to which bead of which DNA segment the protein is eventually attached. Time intervals for which no position is indicated correspond to those periods where the protein is diffusing in the buffer. It was assumed that the protein is attached to bead k of DNA segment j if* $\left\| \mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}} \right\| \leq \sigma$.

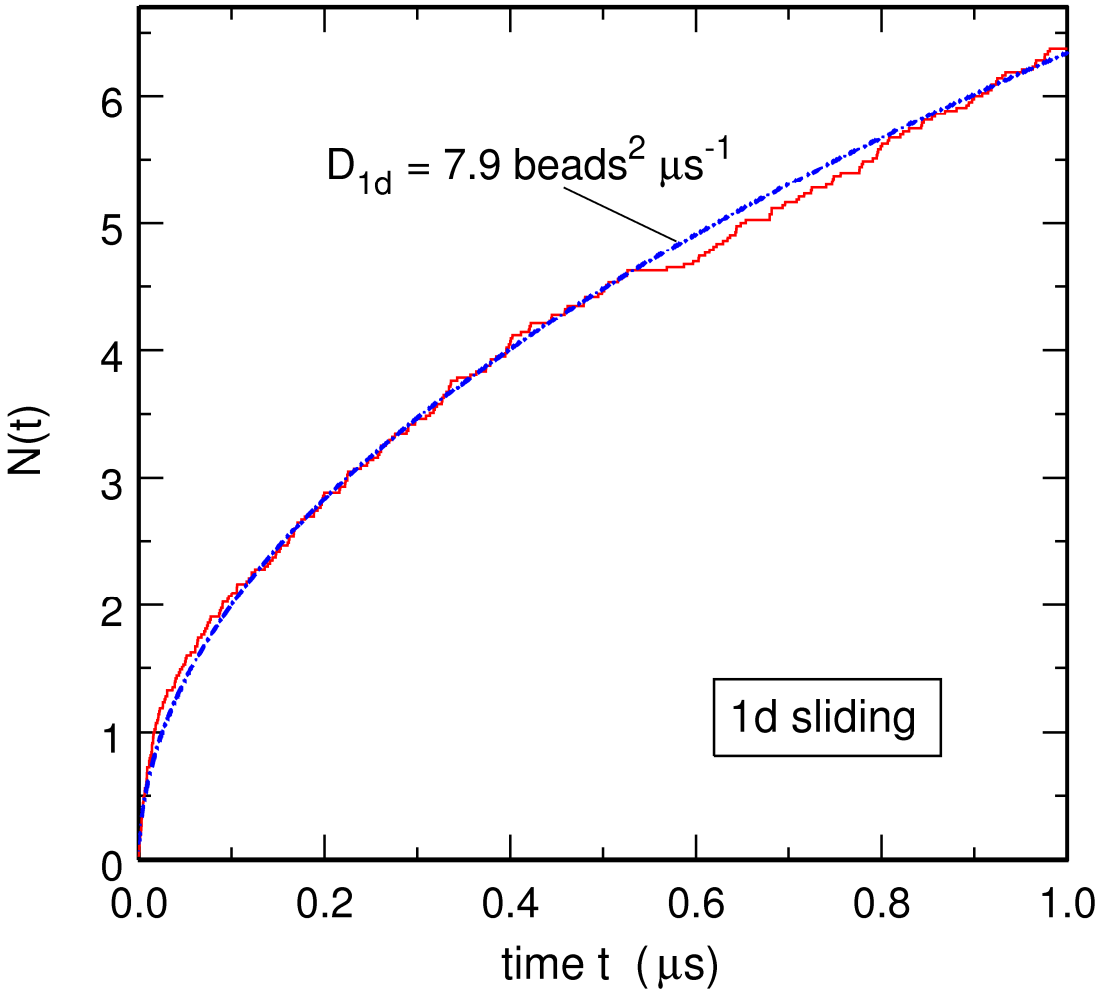

Figure 6.9. *Evolution of the number* $N(t)$ *of different DNA beads visited by the protein during 1D sliding. Calculations were performed with 2000 DNA beads and the ratio* $e_{\text{prot}} / e_{\text{DNA}} = 1$. $N(t)$ *was averaged over 43 sliding events with the following properties : (i) each sliding event lasted more than 1 µs, (ii) the protein did not separate from the DNA segment by more than* σ *during more than 0.07 µs, (iii) the protein bead did not reach one of the extremities of the DNA segment. The dot-dashed line corresponds to a diffusion coefficient* $D_{1D} = 7.9$ *beads*^{2} *µs*$^{-1}$.

follows the dot-dashed line, which represents the evolution of $N(t) = l_0^{-1}\sqrt{16\pi^{-1}D_{1D}\,t}$ with a diffusion coefficient $D_{1D} = 7.9$ beads^2 μs^{-1}, or $D_{1D} \approx 1800$ bp^2 μs^{-1}. The experimental values for the 1D diffusion coefficient of DNA binding proteins are close to 5 (base pairs)^2 μs^{-1} [99,102]. Since one bead represents 15 base pairs, this implies that the model predicts a velocity for sliding, which is about one to two orders of magnitude too large. This is due firstly to the fact that real protein sliding follows the helical path of the DNA chain and is often accompanied by geometrical rearrangements of the DNA sequence, two points that are completely neglected in this model. Moreover, in addition to the $E_e^{(P)}$ electrostatic interaction, the protein and the DNA sequence interact through several hydrogen bonds when the protein is sufficiently close to the sequence. This point is crucial for *specific* DNA-protein interaction (that is, sequence recognition) [13,134-138] but is again completely neglected in the proposed model for *non-specific* DNA-protein interactions. The situation changes somewhat in the case where the protein is described using a more precise model or if it has higher charges, but I will come back to this point later.

Examination of figure 6.6 indicates that the amount of time ρ_{1D}, during which the protein is attached to a DNA segment and experiences sliding, is a monotonically increasing function of the charge on the protein e_{prot}. In contrast, the number $N(t)$ of different DNA beads visited by the protein after a certain amount of time *t* is *not* a monotonic function of e_{prot}, and therefore of ρ_{1D}, as can be checked in figures 6.10 and 6.11. These figures display the evolution of $N(t)$ for the repulsive interaction potential of figure 6.2 and seven values of e_{prot}/e_{DNA} ranging from 0.3 to 5. In figure 6.10, it is assumed that the protein is attached to bead *k* of DNA segment *j* if $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{prot}\right\| \le \sigma$, while the corresponding criterion in figure 6.11 is $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{prot}\right\| \le 1.5\sigma$. It is seen in both figures that $N(t)$ increases with the charge until $e_{prot}/e_{DNA} \approx 1$, then remains nearly constant up to $e_{prot}/e_{DNA} \approx 3$, before decreasing again. The reason for this sharp decrease at large values of e_{prot}/e_{DNA} can be understood from the inspection of figure 6.12, which shows the average number of DNA beads that are simultaneously attached to the protein when it is not moving freely in the buffer. One observes that the number of DNA beads within 1.5σ of the protein is close to 2 for values of e_{prot}/e_{DNA} smaller or close to 1, which indicates that the protein

forms a triangle with two successive DNA beads belonging to the same segment and separated by about $l_0 = 5$ nm. The number of DNA beads within 1.5σ of the protein increases however rapidly for larger values of $e_{\text{prot}} / e_{\text{DNA}}$, because the charge of the protein bead is sufficient to attract several DNA segments, which form a cage around it. The protein visits the DNA beads forming the cage in a short amount of time, but the slope of $N(t)$ then decreases as the protein experiences difficulties to escape the cage and visit other segments. This cage effect is strong enough for the $N(t)$ curve for $e_{\text{prot}} / e_{\text{DNA}} = 5$ to be lower than that for the repulsive potential when the $\left\| \mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}} \right\| \leq 1.5\sigma$ criterion is considered (see figure 6.11).

However, the crucial point is certainly that figures 6.10 and 6.11 show that, for this model, the combination of 1D sliding and 3D motion leads, in a certain range of the $e_{\text{prot}} / e_{\text{DNA}}$ ratio, to faster DNA sampling than pure 3D diffusion. I now assume that nature selects the fastest process and focus on the properties of the system with $e_{\text{prot}} / e_{\text{DNA}} = 1$. Figure 6.13 shows the time evolution of $N(t)$ for systems with $e_{\text{prot}} / e_{\text{DNA}} = 1$ and increasing numbers of DNA beads, namely $mn = 990$, 2000 and 4000. As expected, the three curves coincide at short times, that is, when $N(t) << mn$. Each curve then successively displays saturation as $N(t)$ approaches *mn*. All these curves however follow the law of equation (6.20) with the same rate $\kappa = 1.84\ \mu\text{s}^{-1}$, as can be seen in figure 6.14. This is rather interesting since it indicates that the observed behaviour is independent of the size of the cell and can reasonably be extrapolated to larger cell sizes. In addition, this law implies that even in the case where the protein and the DNA beads attract each other, the global motion of the protein is likely to remain diffusive-like, since $N(t)$ follows at short times the linear law predicted by the formula for the volume of the Wiener sausage (I will come back later in more detail to this important point). Also, the search process is about two times faster than for the case where there are only pure repulsive interactions (for this latter case I obtained a rate $\kappa = 1.09\,\mu\text{s}^{-1}$). This means that facilitated diffusion indeed accelerates this process, but with a factor much smaller than the maximum acceleration predicted by kinetic models. I will present a more detailed comparison between dynamical and kinetic models in chapter 8.

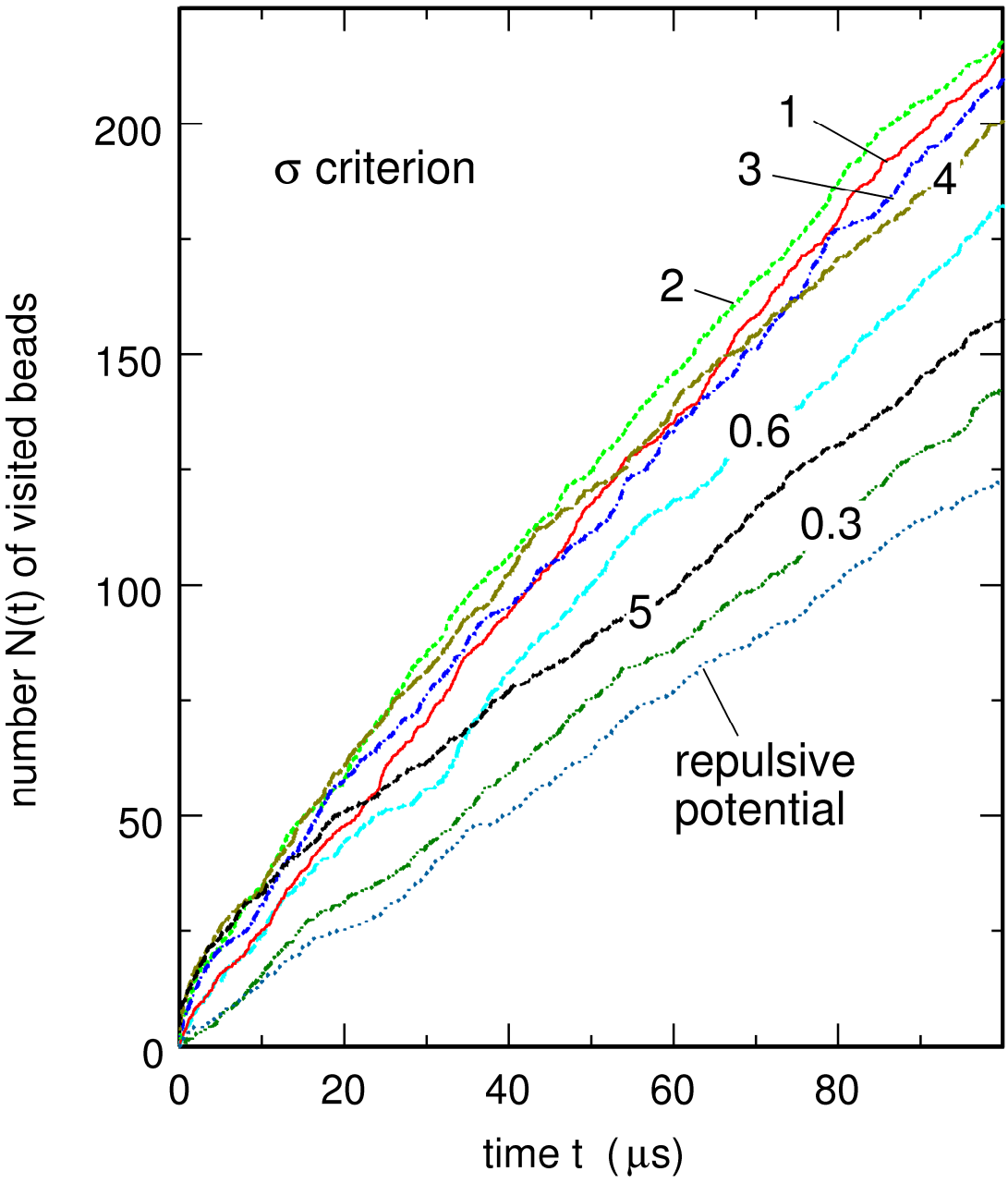

Figure 6.10. *Evolution of the number* $N(t)$ *of different DNA beads visited by the protein, for seven values of* $e_{\text{prot}} / e_{\text{DNA}}$ *ranging from 0.3 to 5 and for the repulsive DNA/protein interaction potential of figure 6.2. Each curve was averaged over 12 different trajectories for the system with 2000 beads. It was assumed that the protein is attached to bead k of DNA segment j if* $\left\| \mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}} \right\| \leq \sigma$.

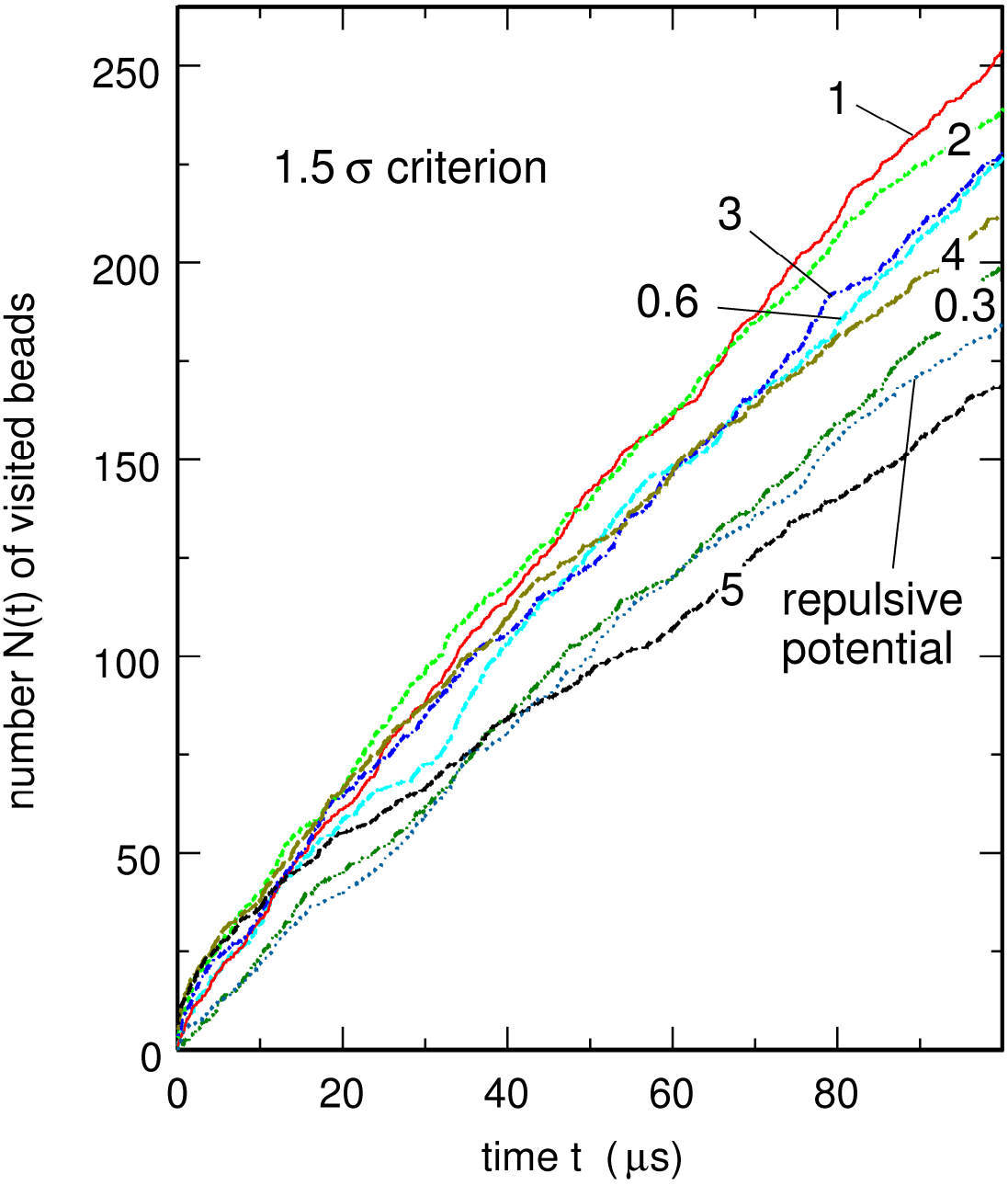

Figure 6.11. *Same as figure 6.10, except that it is considered that the protein is attached to bead k of DNA segment j if* $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \leq 1.5\sigma$ *instead of* $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \leq \sigma$.

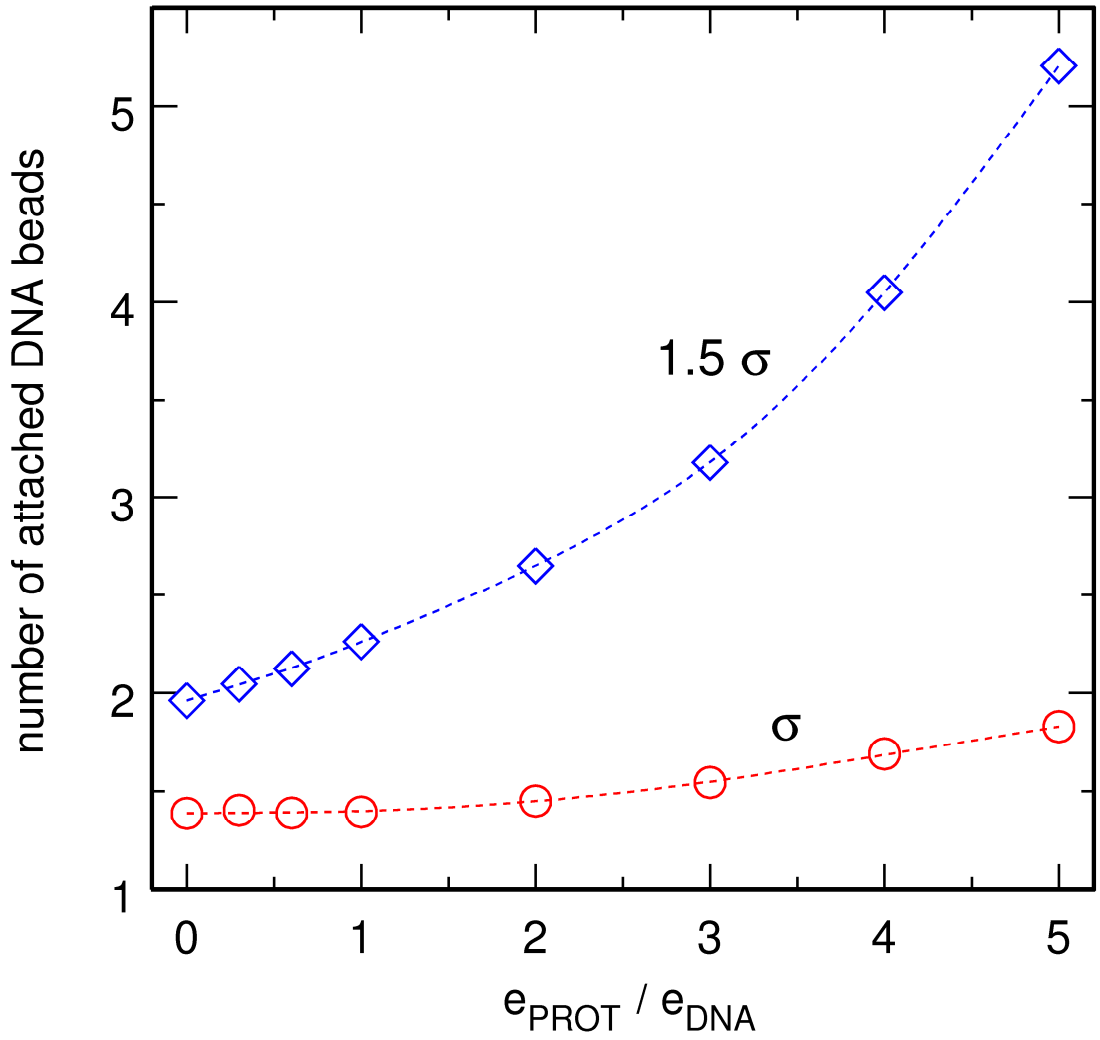

Figure 6.12. *Plot, as a function of the ratio* $e_{\text{prot}} / e_{\text{DNA}}$*, of the average number of DNA beads that are attached to the protein when it does not move freely in the buffer. The abscissa axis actually corresponds to the variation of* e_{prot} *at constant* e_{DNA}*. Circles and lozenges denote results obtained with, respectively, the* $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \le \sigma$ *and* $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \le 1.5\,\sigma$ *criterions for interacting beads. The points at* $e_{\text{prot}} / e_{\text{DNA}} = 0$ *were obtained with the repulsive potential of figure 6.2. Each point was averaged over 12 different trajectories propagated for 100 µs for the system with 2000 beads.*

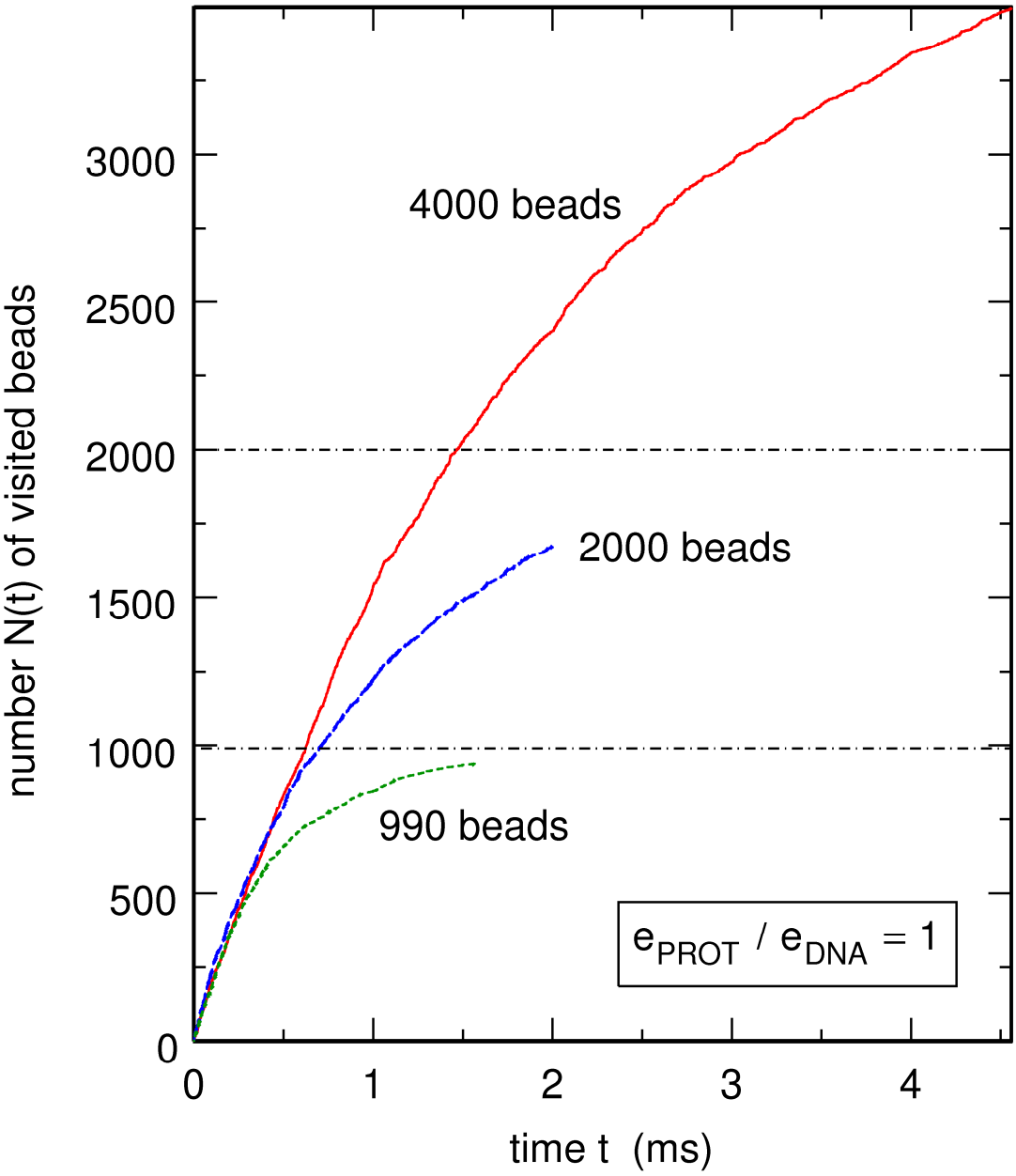

Figure 6.13. *Evolution of* $N(t)$, *the number of different DNA beads visited by the protein at time t, for the system with* $e_{\text{prot}} / e_{\text{DNA}} = 1$ *and 990, 2000 and 4000 DNA beads. It was assumed that the protein is attached to bead k of DNA segment j if* $\left\| \mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}} \right\| \leq \sigma$. *Each curve was averaged over 6 different trajectories.*

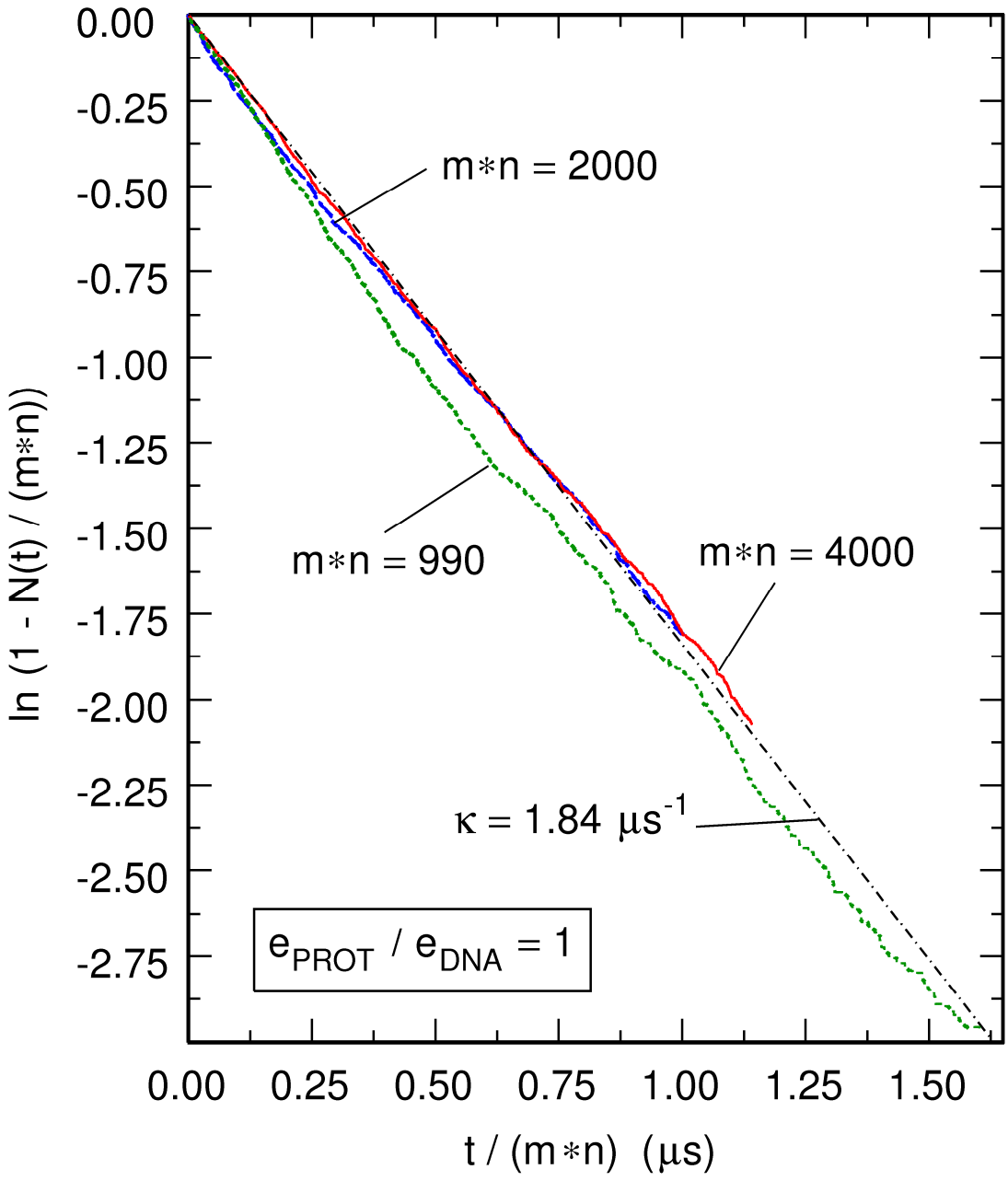

Figure 6.14. *Solid line : plot of* $\ln(1-N(t)/(mn))$ *as a function of* $t/(mn)$ *for the system with* $e_{prot}/e_{DNA}=1$ *and 990, 2000 and 4000 DNA beads (the curves for 2000 and 4000 beads nearly superpose).* $N(t)$ *corresponds to the curves in figure 6.13. The dot-dashed straight line represents the same plot for the expression of* $N(t)$ *in equation (6.20) and a rate* $\kappa=1.84\ \mu s^{-1}$.

6.3. Conclusion

The model described in this chapter is very promising. Despite the fact that it is quite simple, it manages to describe the succession of 1D sliding along the DNA chain and 3D diffusion in the buffer by witch the protein finds its target. However, this model predicts a value of the 1D diffusion coefficient that is too high compared to experimental values. This is a predictable consequence of the approximations that were made. The values obtained here for the sliding length are in good agreement with both experimental results and values predicted by other models. However, it should be noted that, because of the high values that our model predicts for the one-dimensional diffusion coefficient of the protein the predicted duration of a sliding event is necessarily shorter than in the case of experiments. This might imply that it is discussible if the comparison between predicted and measured sliding length is really appropriate.

The roughest approximation certainly concerns the protein, which is described as a single bead with an electric charge e_{prot} placed at its centre. For large values of e_{prot}, this leads to the cage effect discussed previously and to too frequent intersegmental transfers. A better approximation certainly consists in considering the protein as a set of interconnected beads with a certain charge distribution. In the following chapter, I am going to discuss the extent to which the conclusions presented above are affected when the protein is modelled as such a set of interconnected beads.

The model also predicts that the mechanism of facilitated diffusion can indeed accelerate the scanning speed. This acceleration is, however, much more limited than the maximum one predicted by kinetic models. I will come back to this point in chapter 8.

7. Model with 13 beads proteins

The purpose of this chapter is to check the extent to which the conclusions drawn in the preceding chapter are affected when the protein is modelled in a somewhat less crude way.

7.1. The model

The system studied in this chapter is consequently the same as the one described previously but with the protein modelled as a set of thirteen beads connected by springs instead of a single bead. I used two geometries for the protein: “spherical” and “linear” (figure 7.1). The "spherical" protein is obtained by placing 12 beads at the vertices of a regular icosahedron and a thirteenth bead at its centre (21 beads would have been required for a regular dodecahedron). A bond connects the central bead to the 12 other beads, and each bead at a vertex is connected to its five nearest neighbors by a similar bond. The distance between the central bead and those at the vertices is equal to the bead radius $a_{\text{prot}} = 3.5$ nm, so that the radius of the protein at rest is close to 7.0 nm and the distance between two nearest neighbors placed at the vertices is $L_0 = 4a_{\text{prot}} / \sqrt{10 + 2\sqrt{5}} \approx 3.68$ nm. Linear proteins are taken as flexible and extensible chains of 13 beads separated at equilibrium by a distance $a_{\text{prot}} = 3.5$ nm. Because no bending interaction among protein beads is taken into account (see below), "linear" proteins generally assume bent geometries with average end-to-end distances of the order of 17.0 nm. I fixed the number of beads of linear proteins to 13, although they are too large compared to real proteins, for the sake of an easier comparison with spherical proteins.

All beads, except for that at the center of the spherical protein, are assigned electrostatic charges e_p placed at their centers (however, electrostatic interactions between protein beads are neglected, see below). I considered several protein charge distributions, namely (i) uniform distributions with increasing total charge $e_{\text{prot}} = \sum_p e_p$, (ii) gradients of charges with fixed total charge e_{prot} and increasing values of the maximum charge $e_{\max}$, (iii) gradients of charges with fixed maximum charge $e_{\max}$ and increasing total charge e_{prot}, and (iv) random distributions. For spherical proteins, gradient distributions are based on sets of four equally spaced charge values $e_{\max} - k\Delta$, where *k* varies from 0 to 3 and $\Delta = \left(12e_{\max} - e_{\text{prot}}\right)/18$. Charges $e_{\max}$ and $e_{\max} - 3\Delta$

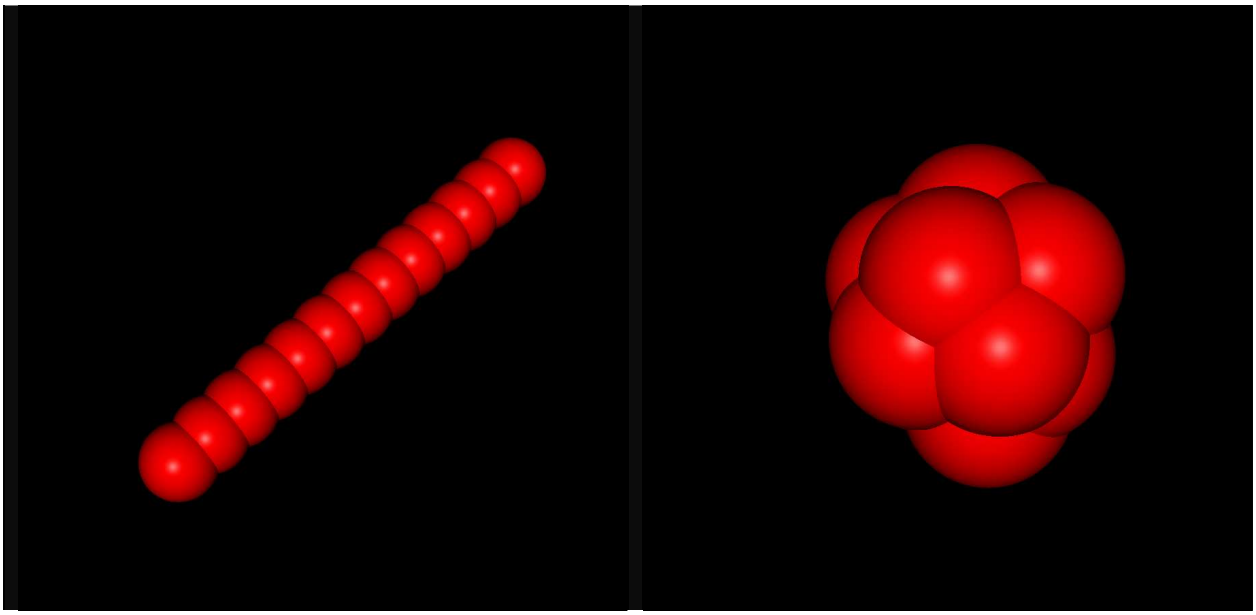

Figure 7.1. *Schematic representations of the two protein models used in chapter 7.*

are carried by two beads placed at opposite vertices of the icosahedron, while the five beads closest to the bead with charge $e_{\max}$ carry a charge $e_{\max} - \Delta$ and the five beads closest to the bead with charge $e_{\max} - 3\Delta$ carry a charge $e_{\max} - 2\Delta$. For linear proteins, instead of a single bead with charge $e_{\max} - 3\Delta$, I placed the charge $(e_{\max} - 3\Delta)/2$ at the centres of two beads, in order to compensate for the fact that the bead placed at the centre of the icosahedron is not charged. For most cases, I increased the total charge e_{prot}, or the maximum charge $e_{\max}$, up to $-5e_{\text{DNA}}$. At last, except otherwise specified, the results presented below were obtained by considering that there is a single bead of the protein that has the ability to connect to DNA and therefore plays the role of a search site. Also, in most cases, and unless otherwise specified, this bead had the highest positive charge $e_{\max}$, but I also ran simulations were this was no longer the case. I have also checked that the results remain essentially the same if one instead assumes that all beads of the protein are able to connect to the DNA chain (figure 7.2).

The potential energy of the system is:

$$E_{\text{pot}} = V_{\text{DNA}} + V_{\text{prot}} + V_{\text{DNA/prot}} + V_{\text{wall}} \tag{7.1}$$

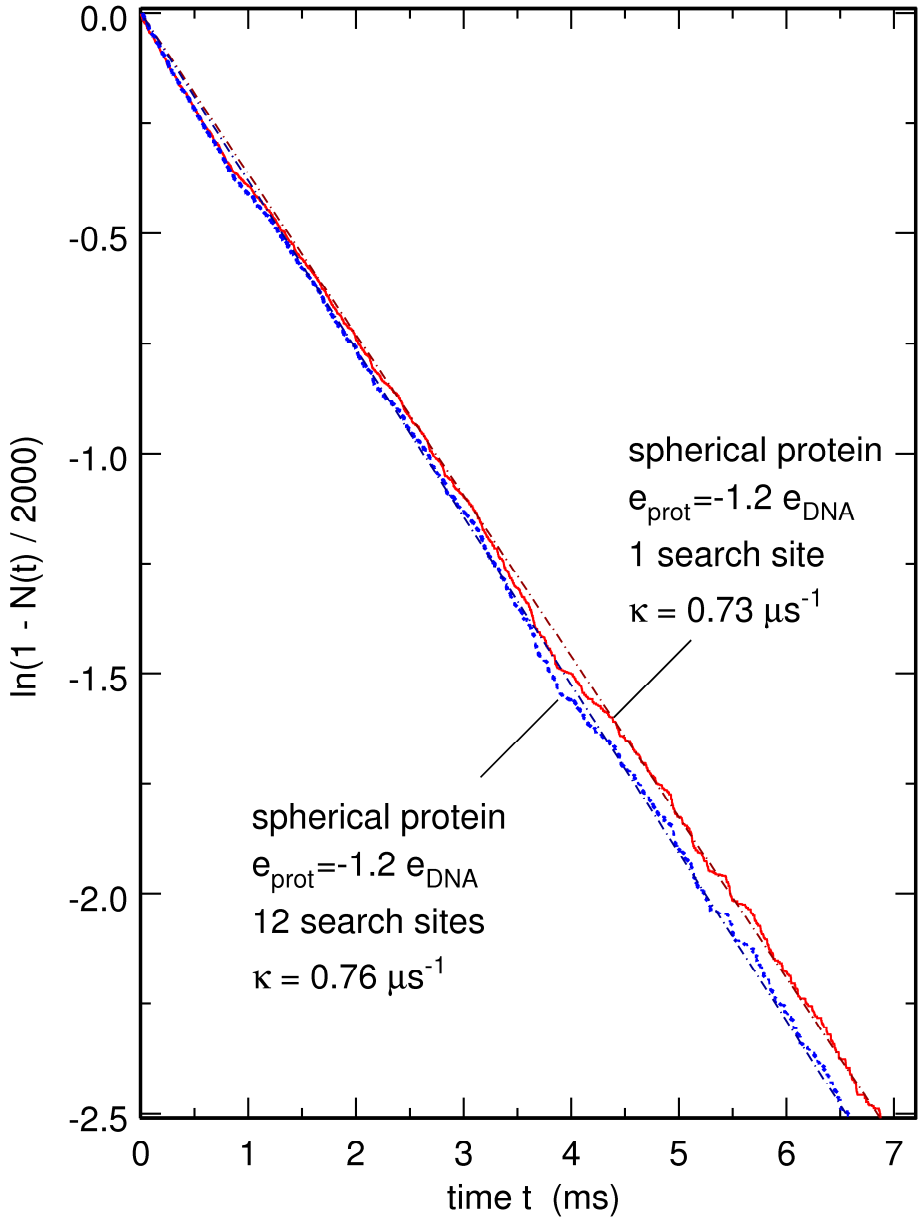

Figure 7.2. *Time evolution of the logarithm of* $1-N(t)/2000$ *, the portion of DNA beads not yet visited by the protein search site, for a spherical protein with a gradient distribution of charges with total charge* $e_{\text{prot}} = -1.2e_{\text{DNA}}$ *and maximum positive charge* $e_{\max} = -0.8e_{\text{DNA}}$ *, for the cases where it has one search site (in red) and twelve search sites (in blue). It was considered that protein bead p is attached to bead k of DNA segment j if* $\left\|\mathbf{r}_{j,k} - \mathbf{R}_p\right\| \le \sigma$.

When comparing to equation (6.1), the extra term, V_{prot}, describes the interaction between the protein beads. The beads that compose the protein interact with each other only by means of harmonic stretching potentials. More precisely, for linear proteins the potential is:

$$V_{\text{prot}} = \frac{1}{2} C \frac{k_B T}{a_{\text{prot}}^2} \sum_{j=0}^{11} \left(L_{j,j+1} - a_{\text{prot}} \right)^2 \tag{7.2}$$

In equation (7.2), the 13 beads are labeled from j=0 to j=12 and $L_{j,j+1} = \left\| \mathbf{R}_j - \mathbf{R}_{j+1} \right\|$ denotes the distance between two successive beads ($\mathbf{R}_j$ is the position of bead j). A distance a_{prot} separates two neighbouring beads at equilibrium. For spherical proteins, I instead assumed that:

$$V_{\text{prot}} = \frac{1}{2} C \frac{k_B T}{a_{\text{prot}}^2} \sum_{j=1}^{12} \left(L_{0,j} - a_{\text{prot}} \right)^2 + \frac{1}{2} C \frac{k_B T}{L_0^2} \sum_{j=1}^{12} \sum_{\substack{k \in V_1(j) \\ k>j}} \left(L_{j,k} - L_0 \right)^2 \tag{7.3}$$

In equation (7.3), index 0 refers to the bead located at the center of the icosahedron and indices 1 to 12 to those placed at the vertices, $L_{j,k} = \left\| \mathbf{R}_j - \mathbf{R}_k \right\|$ denotes the distance between protein beads j and k, and $k \in V_1(j)$ means that the sum runs over the five beads k that are the nearest neighbours of bead j at equilibrium. At equilibrium, the central bead is separated by a_{prot} from the beads placed at the vertices of the icosahedron, while two neighbouring beads located at vertices are separated by L_0. As for the DNA elastic constant h, all the results shown below were obtained, unless otherwise specified, with a constant C in equations (7.2) and (7.3) equal to C=100 in order to get very small displacements of the average bond length without precluding the use of sufficiently large time steps. Still, I also ran simulations where C was varied between 5 and 225 to study how the deformability of proteins affects facilitated diffusion.

The terms for V_{DNA} are the same as in equation (6.2), while V_{wall} becomes

$$V_{\text{wall}} = k_B T \sum_{j=1}^{m} \sum_{k=1}^{n} f\left(\left\| \mathbf{r}_{j,k} \right\| \right) + 10\, k_B T \sum_{j=0}^{12} f\left(\left\| \mathbf{R}_j \right\| \right) \tag{7.4}$$

and $V_{\text{DNA/prot}}$ is modified as follows:

$$V_{\text{DNA/prot}} = \sum_{p=0}^{12} \left(E_{\text{e}}^{(p)} + E_{\text{ev}}^{(p)} \right)$$

$$E_{\text{e}}^{(p)} = \frac{e_{\text{DNA}} e_p}{4\pi\varepsilon} \sum_{j=1}^{m} \sum_{k=1}^{n} \frac{\exp\left(-\frac{1}{r_D} \left\| \mathbf{r}_{j,k} - \mathbf{R}_p \right\| \right)}{\left\| \mathbf{r}_{j,k} - \mathbf{R}_p \right\|}$$

$$E_{\mathrm{ev}}^{(p)} = 1.86\, k_{\mathrm{B}} T \left| \frac{e_p}{e_{\mathrm{DNA}}} \right| \sum_{j=1}^{m} \sum_{k=1}^{n} F\left(\left\| \mathbf{r}_{j,k} - \mathbf{R}_p \right\| \right) \quad , \tag{7.5}$$

where F is given by equation (6.10). In equation (7.5), the charges are taken as signed quantities, while they were considered as positive quantities in equation (6.9). This is the reason why the minus sign in the expression of $E_e^{(p)}$ disappeared. It is important to emphasize again that, when the charges placed at the centre of the DNA and protein beads have opposite signs, the interaction between the two beads must be minimum at some value close to $\sigma = a_{\mathrm{DNA}} + a_{\mathrm{prot}} = 5.28$ nm, i.e. the sum of the radii of the DNA and protein beads, in order for 1D sliding to take place. The expression for $E_{\mathrm{ev}}^{(p)}$ in equation (7.5) insures that this is indeed the case and that the position of the minimum does not depend on the charge e_p. It should however be mentioned that another change in the model is that now the interaction potential is minimum not when the centers of the two beads are separated by σ, as in chapter 6, but rather when this distance is equal to σ+0.5 nm (this was achieved by introducing the factor 1.86 in the expression of $E_{\mathrm{ev}}^{(p)}$). The minimum of the potential well was shifted by this small amount in order to better agree with recent theoretical models [139] and experimental results for complexes of EcoRV [140] and the Skn1 and Sap1 proteins [141].

For solving the equations of motion of the complete system, I used the same algorithm as in chapter 6, including hydrodynamic interactions between all the beads of the protein. All the calculations were performed using the system with 2000 DNA beads and a time step of 100 ps.

7.2. Results

I investigated a large number of different spherical and linear 13 beads protein models and found that, for all of the cases, $N(t)$ follows the law given in equation (6.20) for single bead proteins, that is:

$$\frac{N(t)}{mn} = 1 - \exp(-\kappa \frac{t}{mn}) \tag{7.6}$$

where $mn = 2000$ is the total number of DNA beads.

This is illustrated in figure 7.3, which shows the time evolution of $\log(1 - N(t)/(mn))$ for selected linear and spherical proteins with uniform and gradient distributions of charges. It is seen

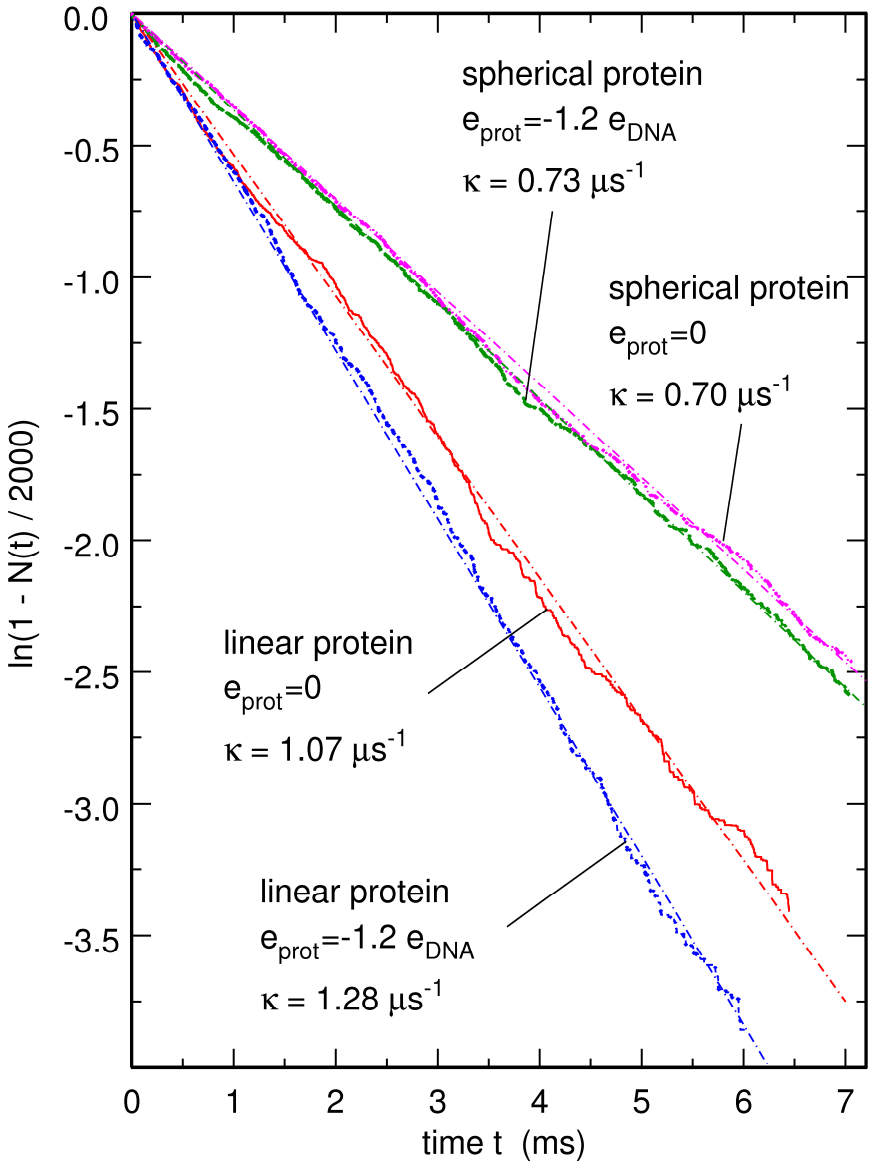

Figure 7.3. *Time evolution of the logarithm of* $1-N(t)/2000$, *the portion of DNA beads not yet visited by the protein search site, for (a) linear proteins with a gradient distribution of charges with total charge* $e_{\text{prot}}=0$ *and maximum charge* $e_{\max}=-0.8e_{\text{DNA}}$ *(solid line), (b) linear proteins with a gradient distribution of charges with total charge* $e_{\text{prot}}=-1.2e_{\text{DNA}}$ *and maximum charge* $e_{\max}=-1.2e_{\text{DNA}}$ *(short dashes), (c) spherical proteins with a gradient distribution of charges with total charge* $e_{\text{prot}}=0$ *and maximum charge* $e_{\max}=-1.5e_{\text{DNA}}$ *(dot-dot-dot-dashes), and (d) spherical proteins with a gradient distribution of charges with total charge* $e_{\text{prot}}=-1.2e_{\text{DNA}}$ *and maximum positive charge* $e_{\max}=-0.8e_{\text{DNA}}$ *(long dashes). For all proteins, the search site was assumed to be the bead with charge* $e_{\max}$. *For the linear proteins, the search site is located at one of the extremities of the chain. It was considered that protein bead p is attached to bead k of DNA segment j if* $\left\|\mathbf{r}_{j,k}-\mathbf{R}_p\right\|\le\sigma$. *The dot-dashed straight lines, which were adjusted against the evolution of* $1-N(t)/2000$ *for each protein, were used to estimate the values of κ.*

that equation (7.6) remains valid for very long times and for values of $N(t)$ very close to the total number of DNA beads. As already mentioned, equation (7.6) reduces to a linear increase at short times. According to the formula for the volume covered by a 3D random walker, this suggests that, as for single bead proteins, the global motion of 13 beads proteins is essentially diffusive-like. Figure 7.3 also points towards a very general result, namely that $N(t)$ increases significantly more rapidly for linear proteins than for spherical ones (at least as long as the search site is located at one of the extremities of the chain, see below). The rationale for this observation is that, according to equation (6.21), κ increases linearly with D_{3D} and the 3D diffusion coefficient of linear proteins is significantly larger than that of spherical ones. I indeed computed the diffusion coefficient for the protein, D_{3D}, from equation (5.11) by launching simulations that involved only the protein and disregarded both DNA segments and cell boundaries. For C=100, I obtained 0.20×10^{-10} m^2/s for the spherical protein and 0.35×10^{-10} m^2/s for the linear one.

In contrast, it might seem at first sight that 13 beads proteins differ more substantially from single bead ones as far as sliding along DNA is involved. For example, figure 7.4 shows log-log plots of $N(t)$ for long sliding events of spherical proteins with uniform and gradient distributions of charges. It is seen that the time evolution of $N(t)$ approximately corresponds to straight lines in these plots, which implies that $N(t)$ increases as a power of t, that is $N(t) = \alpha\, t^{\beta}$, but the exponent β is now smaller than 1/2. Stated in other words, sliding is here subdiffusive. This is not really surprising, because subdiffusion is often encountered in dense media and has recently been experimentally reported for the global motion of proteins in the cytoplasm or the nucleus [142-144]. By looking more closely at sliding events, it can be noticed that 13 beads proteins spend large amounts of time attached to the same DNA beads and the time intervals during which they actually slide are substantially shorter than for single bead proteins with $e_{prot} = -e_{DNA}$. This is an important observation, because it is well known that large average waiting times between random-walk steps are sufficient to induce subdiffusion (see for example [145]). The reason why waiting times are longer for 13 beads proteins than for single bead ones is that, in this model, sliding is driven uniquely by thermal noise and this process is less efficient for 13 beads proteins than for single bead ones, because part of the energy received from collisions is used to deform proteins instead of being converted into sliding impulsions. It might therefore be the case that small barriers, like those observed in figure 6.3, are sufficient to hinder efficiently

the 1D sliding of the protein along the DNA sequence. Still, it should be mentioned that the average number of beads visited during each sliding event (5 to 10 beads, that is from 75 to 150 base pairs) is in fairly good agreement with experimental results, which lie in the range 30 to 170 base pairs [108,109]. This comparison is subject, of course, to the same remark as in chapter 6.

If the depth of the attractive well between DNA and the protein is smaller than the energy $k_B T$ of thermal noise, then the protein does not spend enough time connected to DNA for actual sliding to take place. On the other hand, if attraction is too strong, then the protein remains attached to the same DNA beads instead of sliding. One therefore expects that waiting times become longer for increasing values of the protein charge e_{prot} and, consequently, that the exponent β decreases. It can be checked in the top plot of figure 7.4 that this is indeed the case. While values of β close to 0.40 were obtained for most of the investigated proteins (see figure 7.4), β was found to decrease down to about 0.20 for uniform charge distributions with $e_{\text{prot}} = -4.8 e_{\text{DNA}}$.

At this point, I however checked that single bead proteins actually behave just like 13 beads ones with this respect. More precisely, I performed simulations with single bead proteins with charge $e_{\text{prot}} = -5 e_{\text{DNA}}$ and obtained $\beta \approx 0.30$. The diffusive character of sliding reported in chapter 6 ($\beta = 0.50$ for $e_{\text{prot}} = -e_{\text{DNA}}$) therefore does not extend to proteins with too large values of e_{prot}.

In the previous chapter, the value of the electrostatic charge placed at the center of the protein bead was increased in order to vary the amount of time ρ_{1D} during which the protein is attached to DNA and check whether certain combinations of 1D and 3D motions lead to faster DNA sampling than pure 3D diffusion. Here I will follow the same general idea, except that, since the protein is now modeled by 13 interconnected beads, instead of a single one, there are several different ways to modify ρ_{1D}.

The most natural way to compare the dynamics of the present model to that of the previous one consists in placing identical electrostatic charges at the centre of the 12 beads located at the vertices of the icosahedron (uniform charge distributions) and letting these charges vary. Results for such spherical proteins with uniform charge distributions are presented in figure 7.5. This figure displays the evolution, as a function of the total protein charge e_{prot}, of three

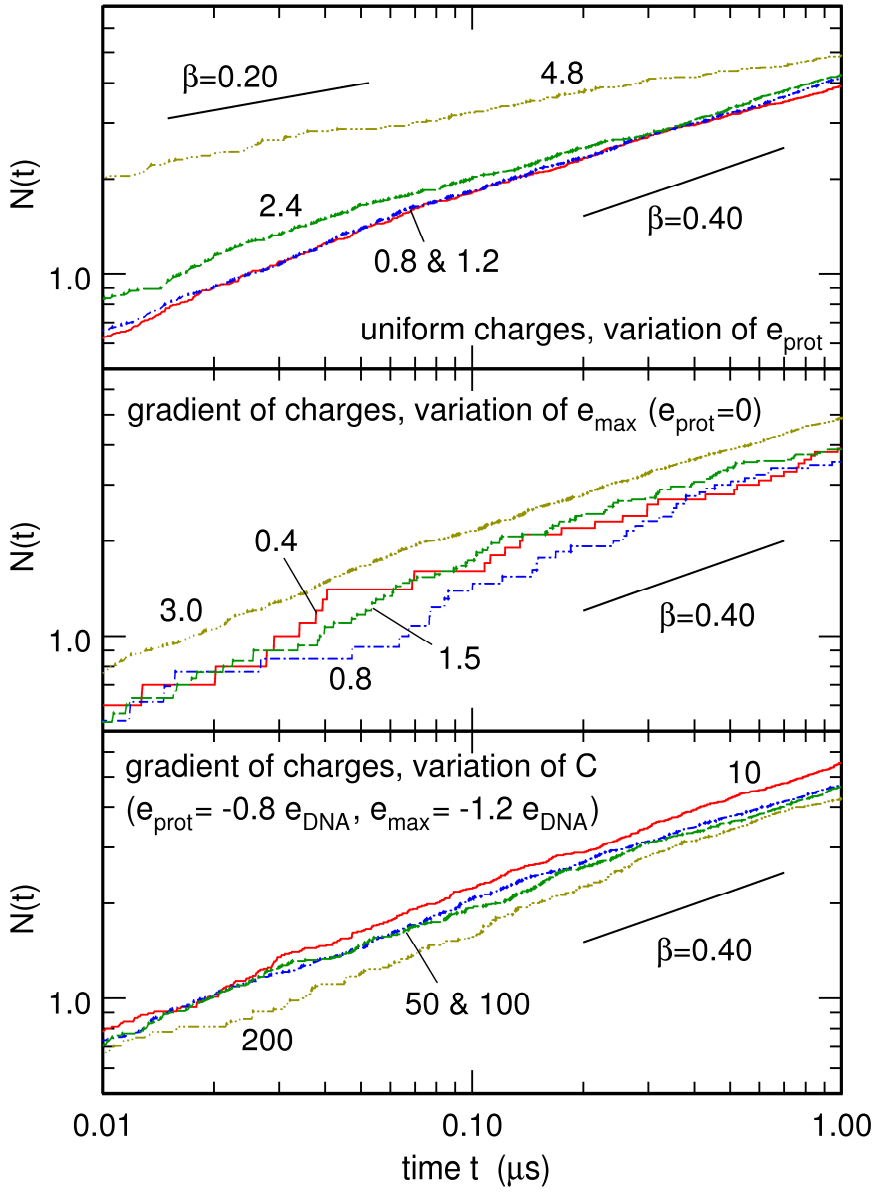

Figure 7.4. *Log-log plots of the time evolution of the number* $N(t)$ *of different DNA beads visited by the protein for spherical proteins with (a) uniform charge distributions and four values of the total charge ranging from* $e_{\text{prot}} = -0.8e_{\text{DNA}}$ *to* $e_{\text{prot}} = -4.8e_{\text{DNA}}$ *(top), (b) gradient distributions of charges with total charge* $e_{\text{prot}} = 0$ *and four values of the maximum charge ranging from* $e_{\text{max}} = -0.4e_{\text{DNA}}$ *to* $e_{\text{max}} = -3e_{\text{DNA}}$ *(middle), and (c) a gradient distribution of charges with total charge* $e_{\text{prot}} = -0.8e_{\text{DNA}}$ *and maximum charge* $e_{\text{max}} = -1.2e_{\text{DNA}}$*, and four values of the elastic constant C ranging from 10 to 200 (bottom). In order to improve the signal/noise ratio, it was assumed for this plot that the protein is attached to bead k of DNA segment j if any of the protein beads (and not a given one) satisfies the condition* $\left\| \mathbf{r}_{j,k} - \mathbf{R}_p \right\| \leq \sigma$*. Each curve was averaged over a number of sliding events that varied between 50 and 200. Each sliding event lasted more than 1 μs, during which the protein neither separated from the DNA segment by more than σ during more than 0.07 μs nor reached one of the extremities of the DNA segment.*

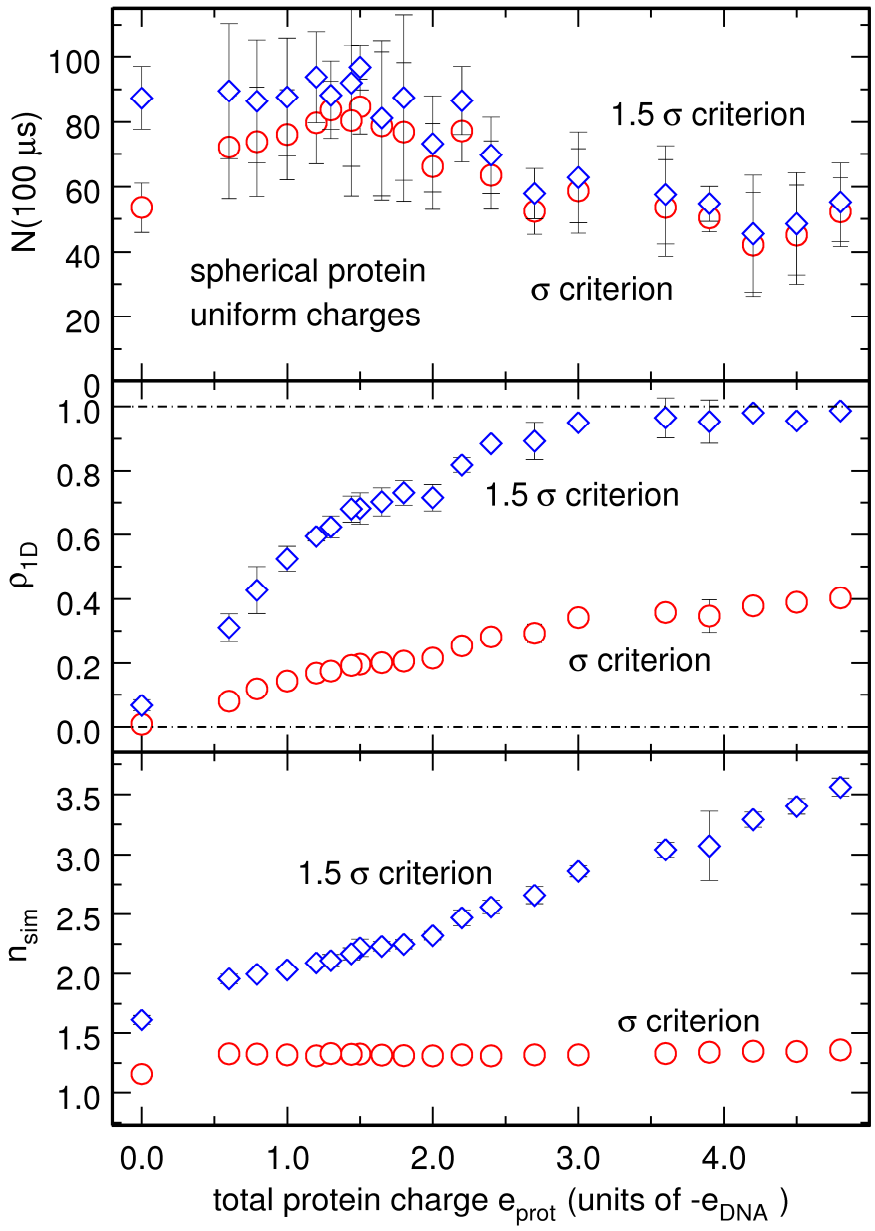

Figure 7.5. *Plot, as a function of the total protein charge* e_{prot}, *of* $N(100\mu s)$ *(top),* ρ_{1D} *(middle), and* n_{sim} *(bottom) for spherical proteins with uniform charge distributions. Circles correspond to results obtained by considering that protein bead p is attached to bead k of DNA segment j if* $\left\|\mathbf{r}_{j,k} - \mathbf{R}_p\right\| \le \sigma$, *while lozenges correspond to the criterion* $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \le 1.5\,\sigma$. *Error bars indicate the standard deviations for the six trajectories over which each point was averaged (note that error bars are masked by circles and lozenges whenever the size of these symbols is larger than the computed standard deviation). Points at* $e_{\text{prot}} = 0$ *denote results obtained with purely repulsive interactions between DNA and the protein.*

quantities, namely $N(100\ \mu s)$, the number of different DNA beads visited by the protein search site after 100 μs (top plot), ρ_{1D}, the portion of time that the protein search site spends attached to a DNA bead (middle plot), and n_{sim}, the average number of DNA beads that are simultaneously attached to the protein search site when it interacts with DNA. Circles correspond to results obtained by considering that protein bead *p* is attached to bead *k* of DNA segment *j* if $\left\|\mathbf{r}_{j,k} - \mathbf{R}_p\right\| \le \sigma$, while lozenges correspond to the criterion $\left\|\mathbf{r}_{j,k} - \mathbf{R}_p\right\| \le 1.5\,\sigma$. Error bars indicate the standard deviations for the six trajectories over which each point was averaged. The points at $e_{prot} = 0$ correspond to purely repulsive DNA-protein interactions, that is, more precisely, when keeping only the repulsive part of the interaction potential with $e_p = -0.1e_{DNA}$. As already mentioned, it can safely be considered that, for repulsive DNA-protein interactions, the motion of the protein inside the cell is rather similar to pure 3D diffusion.

Examination of the middle and bottom plots of figure 7.5 shows that both ρ_{1D} and n_{sim} increase with e_{prot}, like for single bead proteins. Large values of n_{sim} indicate that the protein's charge is sufficiently large for the protein to attract and attach simultaneously to several DNA segments, which form a cage around it. This phenomenon is probably not relevant from the biological point of view, because only a few proteins are known to have more than one "reading head" [13] (the best known example is the *lac* repressor, which has two binding sites [146]). This implies that one should consider only those charge distributions, which are associated with moderate values of n_{sim}, for instance, smaller than 3 for the $1.5\,\sigma$ threshold.

When comparing the top plot of figure 7.5 to figure 6.10, one first notices that $N(t)$ increases more slowly for 13 beads proteins than for single bead ones. For example, for the repulsive potential, the number of DNA beads visited by 13 beads proteins is only about 50% of the number of DNA beads visited by single bead proteins. This is again essentially due to the difference in the values of the 3D diffusion coefficient at 298 K, which is equal to $D_{3D} = 0.70\times10^{-10}$ m^2/s for single beads and to $D_{3D} \approx 0.20\times10^{-10}$ m^2/s for the spherical protein. Nonetheless, the key point is certainly that, as for single bead proteins, there exists a range of values of e_{prot} for which $N(t)$ increases more rapidly than for repulsive DNA-protein interactions. This range extends roughly up to $e_{prot} = -2e_{DNA}$ for uniform charge distributions. It

can be noticed that $N(t)$ is increased at maximum by about 50% compared to the repulsive potential, not so far from the maximum increase close to 70% obtained for single bead proteins.

Needless to say that these conclusions drawn from the dynamics of proteins with uniform charge distributions must be confirmed by results obtained for more complex distributions. I postpone the case of random charge distributions until the next section and focus now on the results obtained for spherical proteins with gradient distributions of charges. For such gradient distributions, I either fixed the value of the maximum protein charge $e_{\max}$ and varied the total charge e_{prot}, or fixed e_{prot} and varied $e_{\max}$. It turns out that the results obtained for these gradient distributions are quite similar to those discussed above, at least as long as e_{prot} and $e_{\max}$ remain moderate. For example, the results for $e_{\max} = -0.8e_{\text{DNA}}$ are shown in figure 7.6 and those for $e_{\text{prot}} = 0$ in figure 7.7. It is seen that, in both cases, $\rho_{1\text{D}}$ increases with increasing charge and $N(100\ \mu\text{s})$ goes through a maximum for values of $\rho_{1\text{D}}$ comprised between 0.3 and 0.7 for the $1.5\,\sigma$ threshold. Moreover, the increase of $N(t)$ relative to the case of purely repulsive interactions between DNA and the protein does not exceed 40%, which again agrees with the results obtained for uniform charge distributions. Things are however noticeably different for larger values of $e_{\max}$ or e_{prot}. For example, I checked that for gradient distributions with $e_{\text{prot}} = -2.4e_{\text{DNA}}$, the total protein charge is sufficiently large for proteins to spend all the time attached to a DNA segment, irrespective of $e_{\max}$ (and consequently of the charge of the search site: I assumed so far that the search site is the protein bead with highest positive charge). As a consequence, $N(100\ \mu\text{s})$ varies little with increasing values of $e_{\max}$.

Conclusion therefore is that, even for rather rigid spherical protein models (remember that *C*=100 for all the results presented above), facilitated diffusion increases DNA sampling speed by about 20 to 50% compared to 3D diffusion, which is even less than the 70% increase observed for single bead proteins. Still, the efficiency of the facilitated diffusion mechanism is again lower for linear proteins, as can be seen in figure 7.8, which shows results obtained for linear proteins with uniform charge distributions (similar results were obtained for gradient distributions with $e_{\text{prot}} = 0$). *C* was also fixed to 100. Since no clear increase of $N(100\ \mu\text{s})$ is observed when the

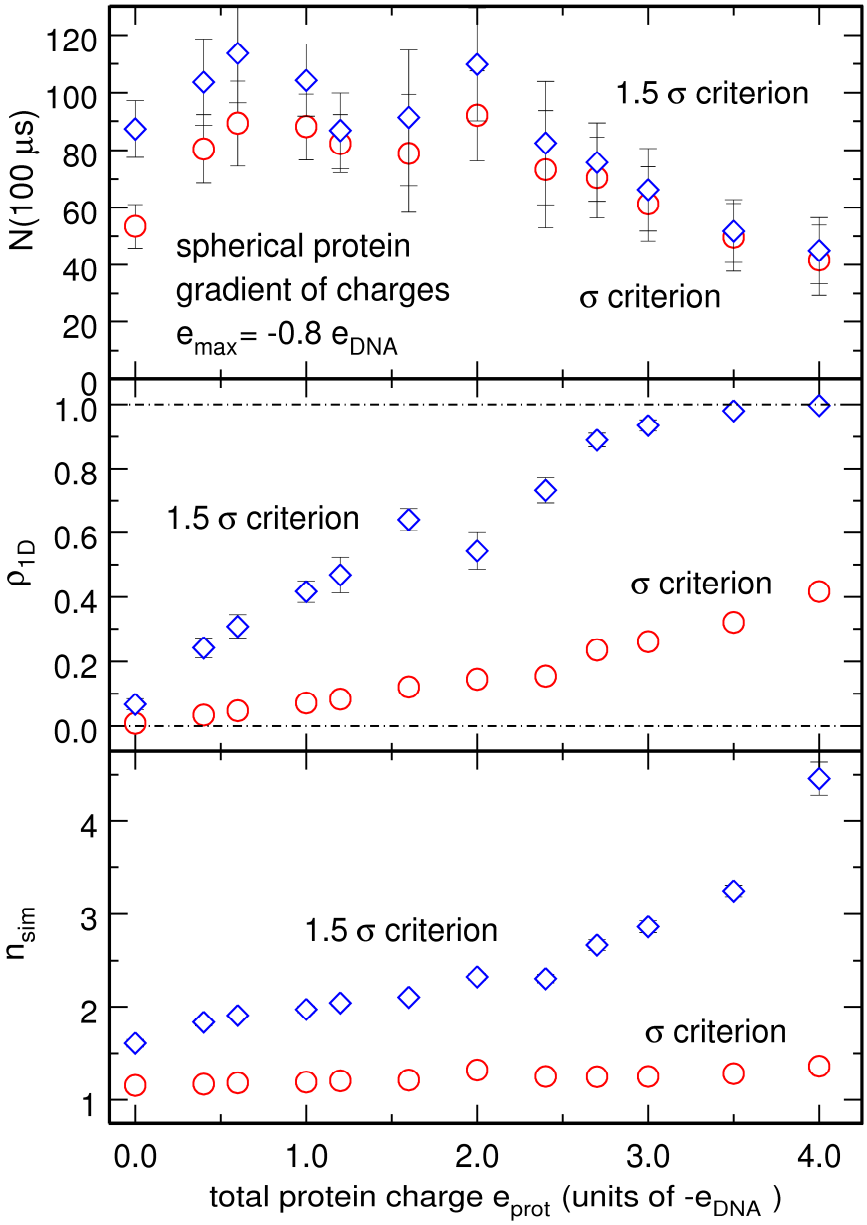

Figure 7.6. *Same as figure 7.5, but for spherical proteins with gradient distributions of charges and maximum positive charge* $e_{\max} = -0.8e_{\mathrm{DNA}}$. *The search site is assumed to be the protein bead with charge* $e_{\max}$.

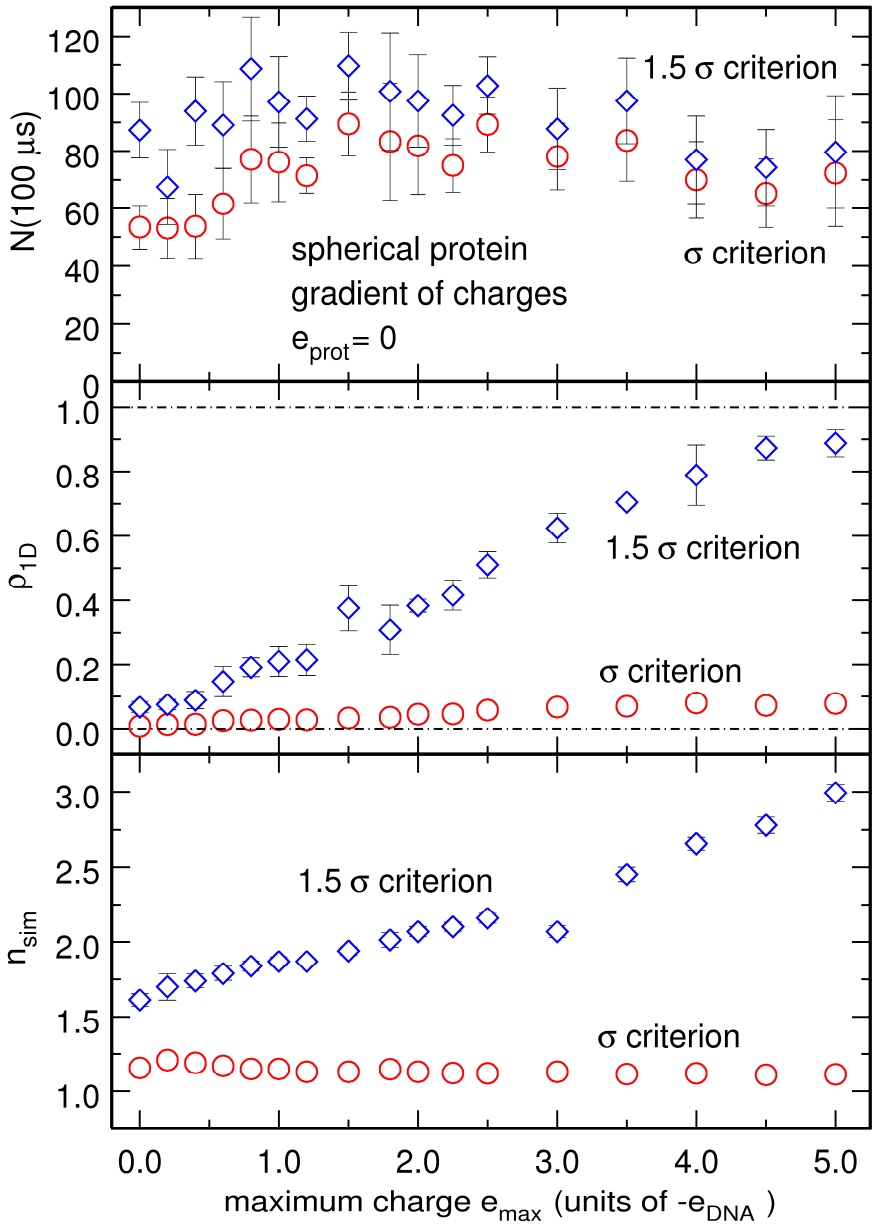

Figure 7.7. *Same as figure 7.5, but for spherical proteins with gradient distributions of charges and total charge* $e_{\text{prot}} = 0$. *The search site is assumed to be the protein bead with charge* e_{max}.

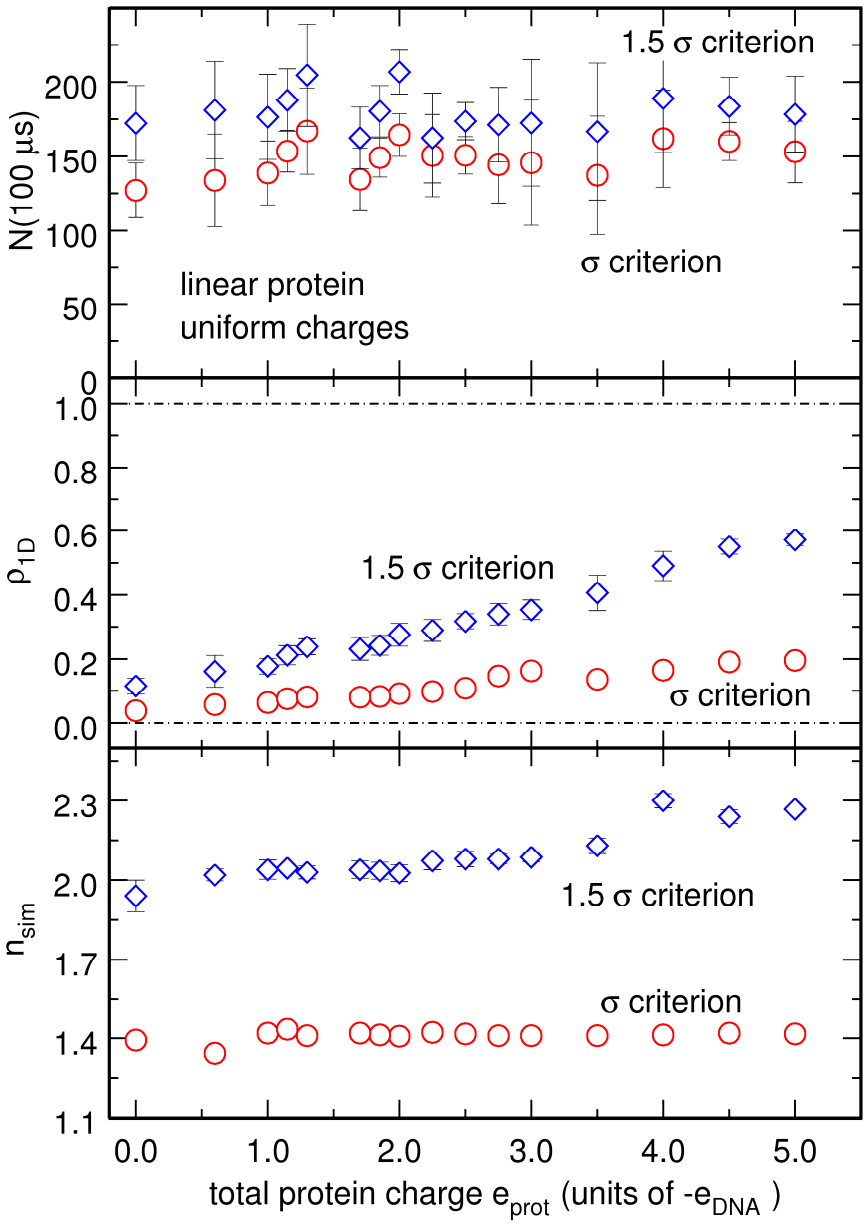

Figure 7.8. *Same as figure 7.5, but for linear proteins with uniform charge distributions. The search site is assumed to be one of the beads located at the extremities of the protein chain.*

total charge is increased from zero, in spite of the fact that ρ_{1D} does increase significantly, it must be admitted that no combination of 1D and 3D motions is more efficient than pure 3D diffusion. This can be understood by noticing that, for identical values of *C*, spherical proteins are much more rigid than linear ones, because each bead at the vertices of the icosahedron is connected to the central bead and to its five nearest neighbours, while each bead of linear proteins is connected to only one or two nearest neighbours. 1D sliding of linear proteins is therefore even less efficient than that of spherical ones.

7.3. Other factors that affect the speed-up of DNA sampling

The purpose of this section is to discuss the effect of several other parameters, namely the value of the elastic constant *C*, the randomness of the charge distribution, and the charge and position of the protein search site, on the speed of DNA sampling.

Let us first consider the effect of the protein elastic constant *C*. The time evolution of the number $N(t)$ of different DNA beads visited by the protein search site for spherical proteins with a gradient distribution of charges with $e_{prot} = -0.8e_{DNA}$ and $e_{max} = -1.2e_{DNA}$ and values of *C* ranging from 10 to 225 is shown in the bottom plot of figure 7.4 for long sliding events and in figure 7.9 for the global (1D+3D) motion. While 1D sliding depends little on *C*, for the global motion $N(t)$ instead decreases significantly and rapidly with *C* for values of *C* comprised between 10 and 100 before remaining nearly constant for larger values of *C*. It can be checked in figure 7.10 that this is essentially due to the evolution of the 3D diffusion coefficient with increasing values of *C*, in agreement with equation (6.21). The top plot indeed shows that D_{3D} decreases from about 0.32×10^{-10} m^2/s for *C*=5 to about 0.20×10^{-10} m^2/s for values of C larger than 100. The average protein radius $\langle L_{0,j} \rangle$ was also computed during these simulations. Results are shown as filled circles in the bottom plot of figure 7.10. It is seen that $\langle L_{0,j} \rangle$ *decreases* with increasing values of *C* in the range *C*=5-100, so that the decrease of D_{3D} in this range is in clear contradiction with equation (6.23). Note that this decrease of $\langle L_{0,j} \rangle$ with increasing *C* agrees with preceding work [130]. It is actually due to hydrodynamic interactions. Indeed, if

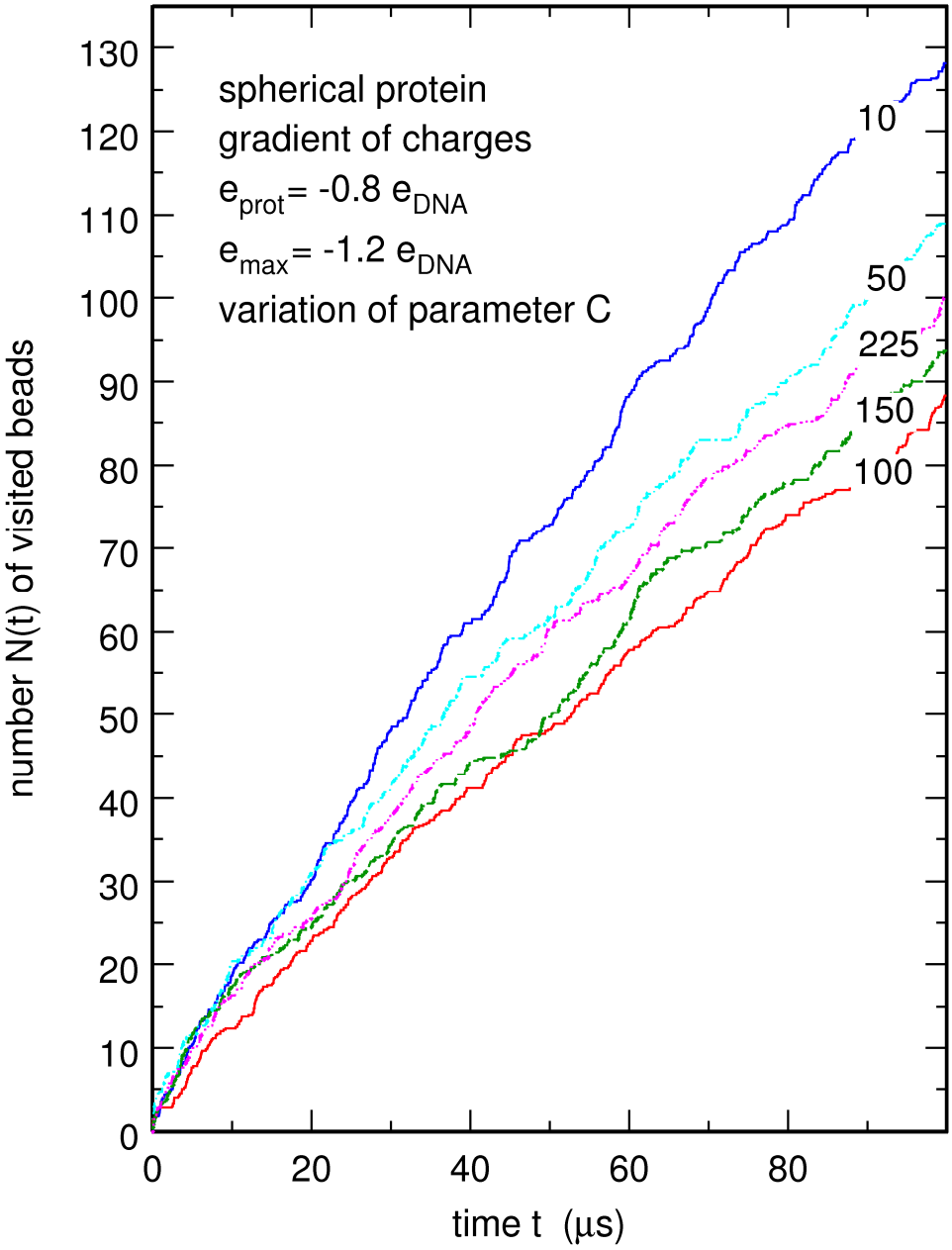

Figure 7.9. *Time evolution of the number* $N(t)$ *of different DNA beads visited by the protein search site for spherical proteins with a gradient distribution of charges with* $e_{\text{prot}} = -0.8e_{\text{DNA}}$ *and* $e_{\text{max}} = -1.2e_{\text{DNA}}$*, and five values of the elastic constant C ranging from 10 to 225. The value of C is indicated for each curve. It was considered that protein bead p is attached to bead k of DNA segment j if* $\left\| \mathbf{r}_{j,k} - \mathbf{R}_p \right\| \leq \sigma$.

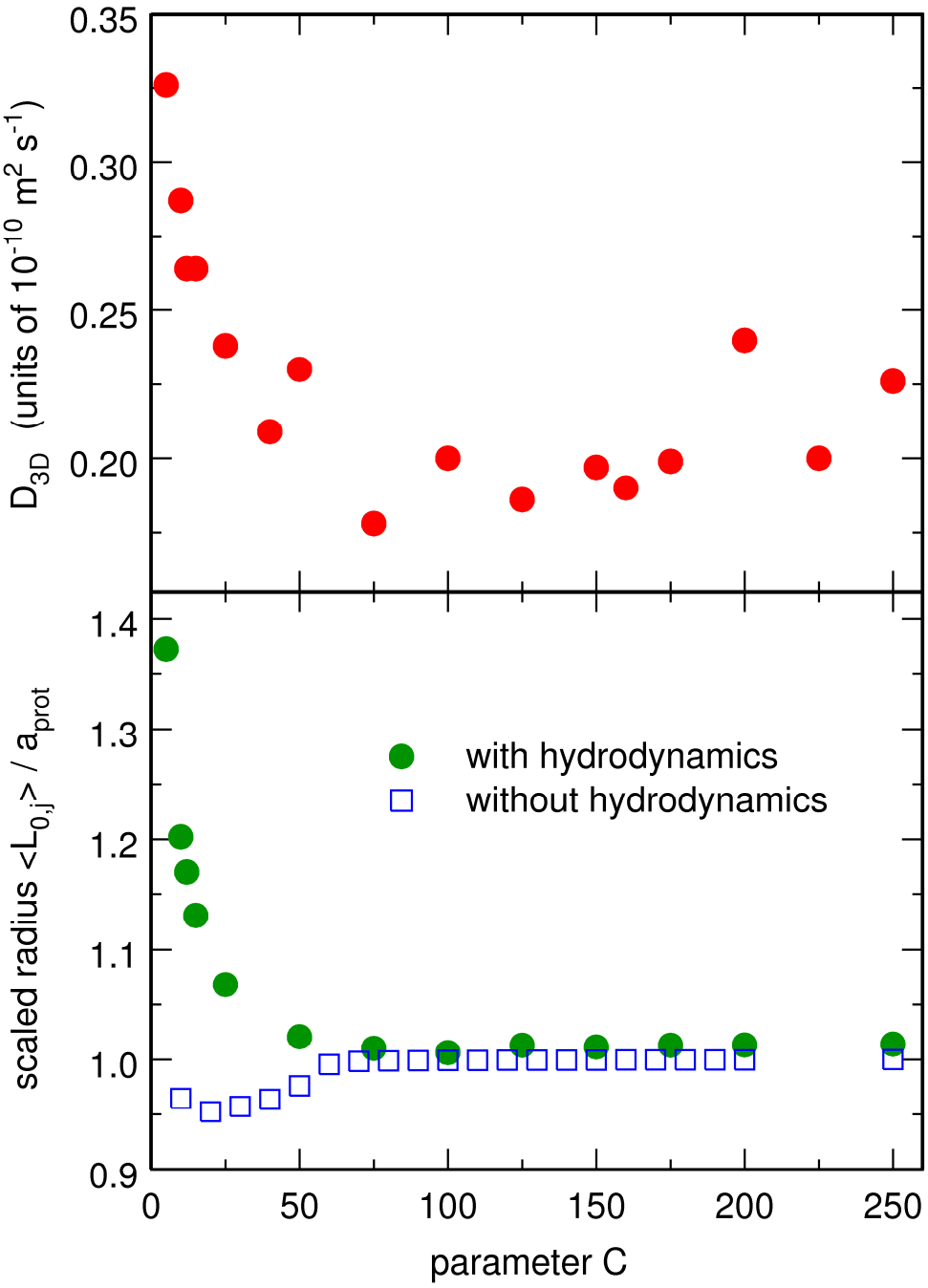

Figure 7.10. *Evolution, as a function of the value of the elastic constant* C*, of (a)* D_{3D}*, the 3D diffusion coefficient at 298 K of spherical (top plot), and (b) the average value of* $L_{0,j}/a_{prot}$ *for these proteins, obtained from simulations with (filled circles) and without (empty squares) hydrodynamic interactions.* $L_{0,j}$ *is the distance between the central bead with index 0 and the bead with index j>0 initially located at one of the vertices of the icosahedron.*

hydrodynamic interactions are neglected, then $\langle L_{0,j} \rangle$ evolves only very little with C in the range C=5-100 (see the empty squares in the bottom plot of figure 7.10). This points out that, for a system where hydrodynamic interactions are expressed by the Rotne-Prager tensor, there is not necessarily a $1/r$ dependence of the translational diffusion coefficient, as given by the Einstein formula in equation (6.23). Actually, the displacement from this equation can be evaluated by means of the Kirkwood-Riseman formula, which gives the translational diffusion coefficient of a chain of beads using pre-averaged values for hydrodynamic forces [147]:

$$D = \frac{k_B T}{6\pi\eta a} \frac{1}{N} \left(1 + \frac{a}{N} \sum_{i \neq j} \left\langle r_{ij}^{-1} \right\rangle \right) \tag{7.7}$$

where $\langle r_{ij}^{-1} \rangle$ is the mean inverse distance between beads i and j averaged over an ensemble of configurations. In conclusion, the deformability of the protein essentially affects the speed of DNA sampling through the associated variations of the diffusion coefficient, much as the shape of the protein that was previously discussed.

Another parameter that might affect the DNA sampling process is the regularity/randomness of the protein charge distribution. While all results presented up to now involved proteins with either uniform or gradient distributions of charges, figure 7.11 indicates how these results are affected when the charges of a gradient distribution are redistributed randomly. More precisely, this figure shows the time evolution of $N(t)$ for spherical proteins with a gradient distribution of charges with $e_{\text{prot}} = -2.4e_{\text{DNA}}$ and $e_{\max} = -1.2e_{\text{DNA}}$, as well as two distributions obtained by random permutations of these charges (but the search site remains the bead with charge $e_{\max}$). It can be checked on this example that the regular and random charge distributions lead essentially to the same behaviour for $N(t)$.

A related question is that of the importance of the charge carried by the search site. At this point it should be remembered that it was assumed in all simulations discussed up to now that the search site is the bead with largest positive charge $e_{\max}$. However, results are not much affected when this condition is released. For example, the time evolution of $N(t)$ for spherical proteins with identical gradient distributions of charges with $e_{\text{prot}} = -1.2e_{\text{DNA}}$ and $e_{\max} = -1.2e_{\text{DNA}}$ but search sites located either on bead 1 (with charge $e_{\max} = -1.2e_{\text{DNA}}$) or bead 2 (with charge $-0.467e_{\text{DNA}}$) are compared in figure 7.12. It is seen that the difference between the two curves is

not significant. By combining the two later observations, it can be surmised that the results should be rather similar for a given set of protein charges, whatever the exact spatial distribution of the charges and the precise charge carried by the search site. It can be checked in figure 7.13 that this is indeed the case. This figure shows the time evolution of $N(t)$ for linear proteins with a gradient distribution of charges with $e_{\text{prot}} = -2.4e_{\text{DNA}}$ and $e_{\max} = -1.2e_{\text{DNA}}$ (solid line), as well as two distributions obtained by random permutations of these charges. The search site is the central (seventh) bead of each chain. It has a charge of $0.13e_{\text{DNA}}$ for the gradient distribution, and charges $-0.53e_{\text{DNA}}$ and $0.40e_{\text{DNA}}$ for the random distributions. In spite of the large differences between these proteins, the evolution of $N(t)$ is essentially similar for the three of them.

Conclusion therefore is that, within the validity of this coarse grained model, the dynamics of DNA sampling is essentially governed by the total charge of the protein or, in the case this charge is small, by the maximum local charge, but that the exact spatial distribution of charges and the precise charge carried by the search site play little role. It can of course not be excluded that this conclusion will be somewhat moderated when the dynamics of finer grained models is investigated.

In contrast, it should be mentioned that a factor that certainly does play an important role is the accessibility of the protein search site. For example, it is clear that, for linear proteins, beads located at the extremities of the chain are more accessible and have a higher probability to interact with DNA than beads located inside the chain, so that one expects DNA sampling by the former ones to be more efficient. This is confirmed by the examination of figure 7.12, which displays the time evolution of $N(t)$ for linear proteins with identical uniform charge distributions with total charge $e_{\text{prot}} = -1.3e_{\text{DNA}}$, but with search sites placed either on bead 1 (extremity) or bead 7 (central bead). It is seen that bead 1 samples DNA at a speed about 50% larger than the central bead. This conclusion obviously agrees with the observation that, in real life, "reading heads" are usually exposed outside the proteins, like the two α helices of the *cro* repressor, which can be inserted in the major or minor grooves of the DNA double helix [13].

7.4. Conclusion

In this chapter, I improved the molecular mechanical model presented in chapter 6 by describing the protein as a set of interconnected beads instead of a single one. Most of the results obtained with this improved model agree with both experimental results and the predictions of the previous model. The new model predicts, like the original one, that DNA sampling proceeds via a succession of 3D motion in the cell, 1D sliding along the DNA sequence, short or long hops between neighbouring or more widely separated sites, and intersegmental transfers. This more detailed description for the protein permitted to show that, within the validity limits of this model, the shape and deformability of proteins essentially affect the speed of DNA sampling trough the associated variations of their diffusion coefficient. Moreover, this model predicts that the sampling speed is governed by the total charge on the protein rather than by that on the search site. Also, this model predicts an acceleration of site targeting due to facilitated diffusion that is even smaller than what was predicted in the previous chapter. Since this result seems to be in contradiction with the predictions of many kinetic models, I will present in the next chapter a detailed comparison between dynamical and kinetic models.

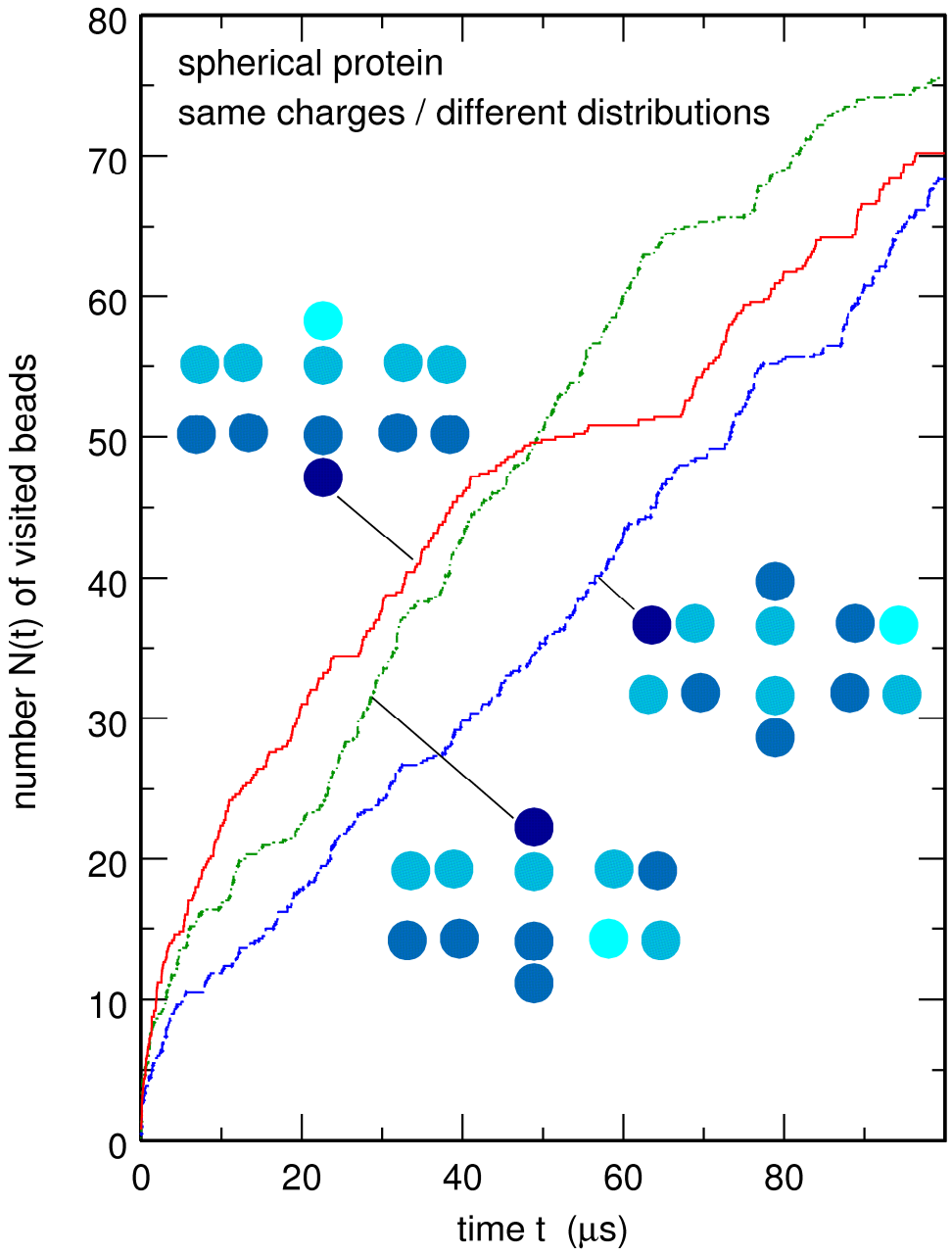

Figure 7.11. *Time evolution of the number* $N(t)$ *of different DNA beads visited by the protein search site for spherical proteins with a gradient distribution of charges with total charge* $e_{\text{prot}} = -2.4e_{\text{DNA}}$ *and maximum charge* $e_{\max} = -1.2e_{\text{DNA}}$ *(solid line), as well as two distributions obtained by random permutations of these charges. Shown in the small inserts are the positions of the charges at equilibrium. The darkest disk corresponds to charge* $e_{\max} = -1.2e_{\text{DNA}}$ *and the brightest one to the maximum negative charge* $0.8e_{\text{DNA}}$*. The search site is the protein bead with charge* $e_{\max}$*. It was considered that protein bead p is attached to bead k of DNA segment j if* $\left\| \mathbf{r}_{j,k} - \mathbf{R}_p \right\| \le \sigma$.

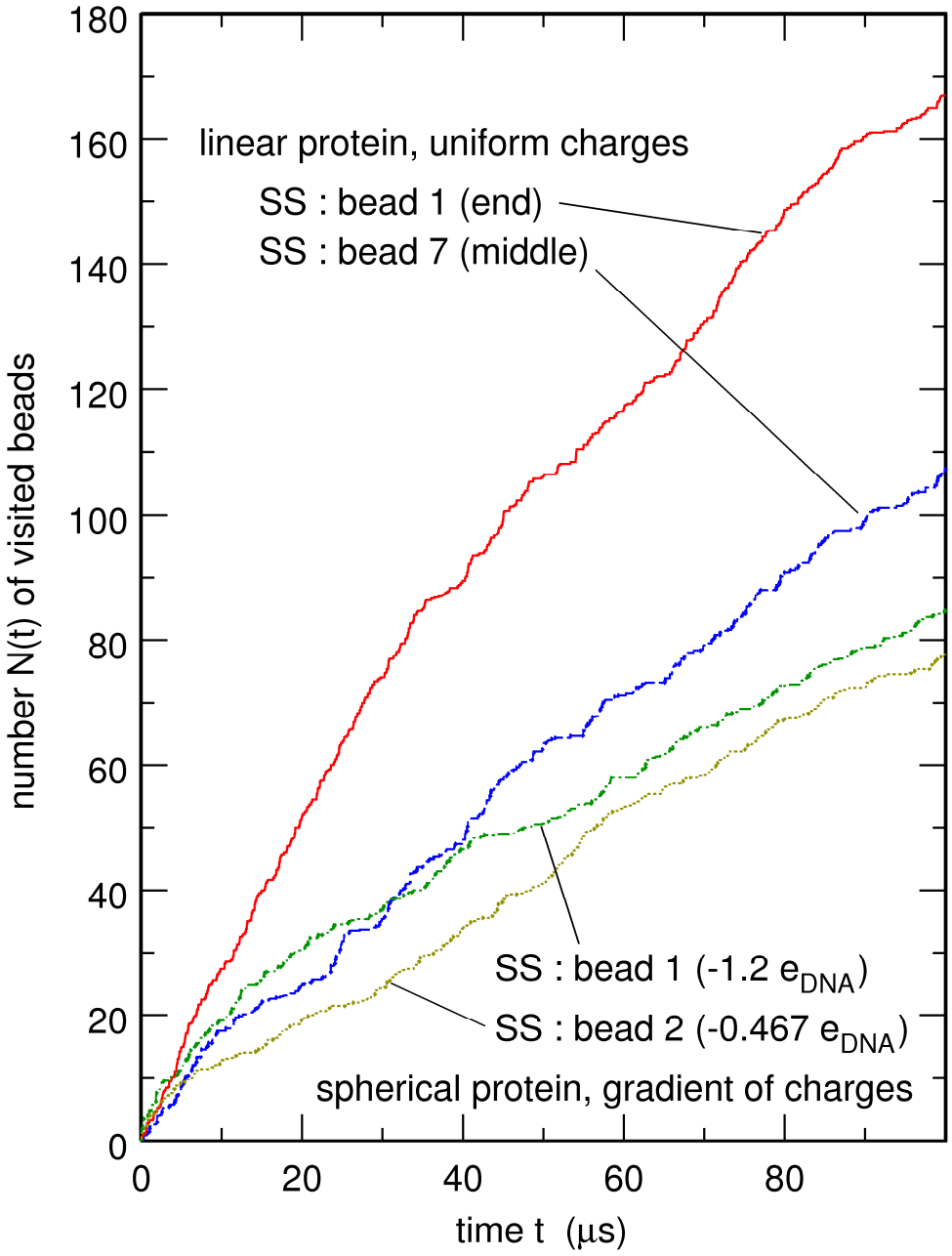

Figure 7.12. *Time evolution of the number* $N(t)$ *of different DNA beads visited by the protein search site for linear proteins with a uniform charge distribution with total charge* $e_{\text{prot}} = -1.3e_{\text{DNA}}$, *as well as for spherical proteins with a gradient distribution of charges with total charge* $e_{\text{prot}} = -1.2e_{\text{DNA}}$ *and maximum charge* $e_{\max} = -1.2e_{\text{DNA}}$. *For the linear proteins, the search site (SS) is assumed to be either the first or the seventh (middle) bead, while for the spherical proteins the SS is assumed to be either bead 1 with charge* $e_{\max} = -1.2e_{\text{DNA}}$ *or bead 2 with charge* $-0.467e_{\text{DNA}}$. *It was considered that protein bead p is attached to bead k of DNA segment j if* $\left\|\mathbf{r}_{j,k} - \mathbf{R}_p\right\| \le \sigma$.

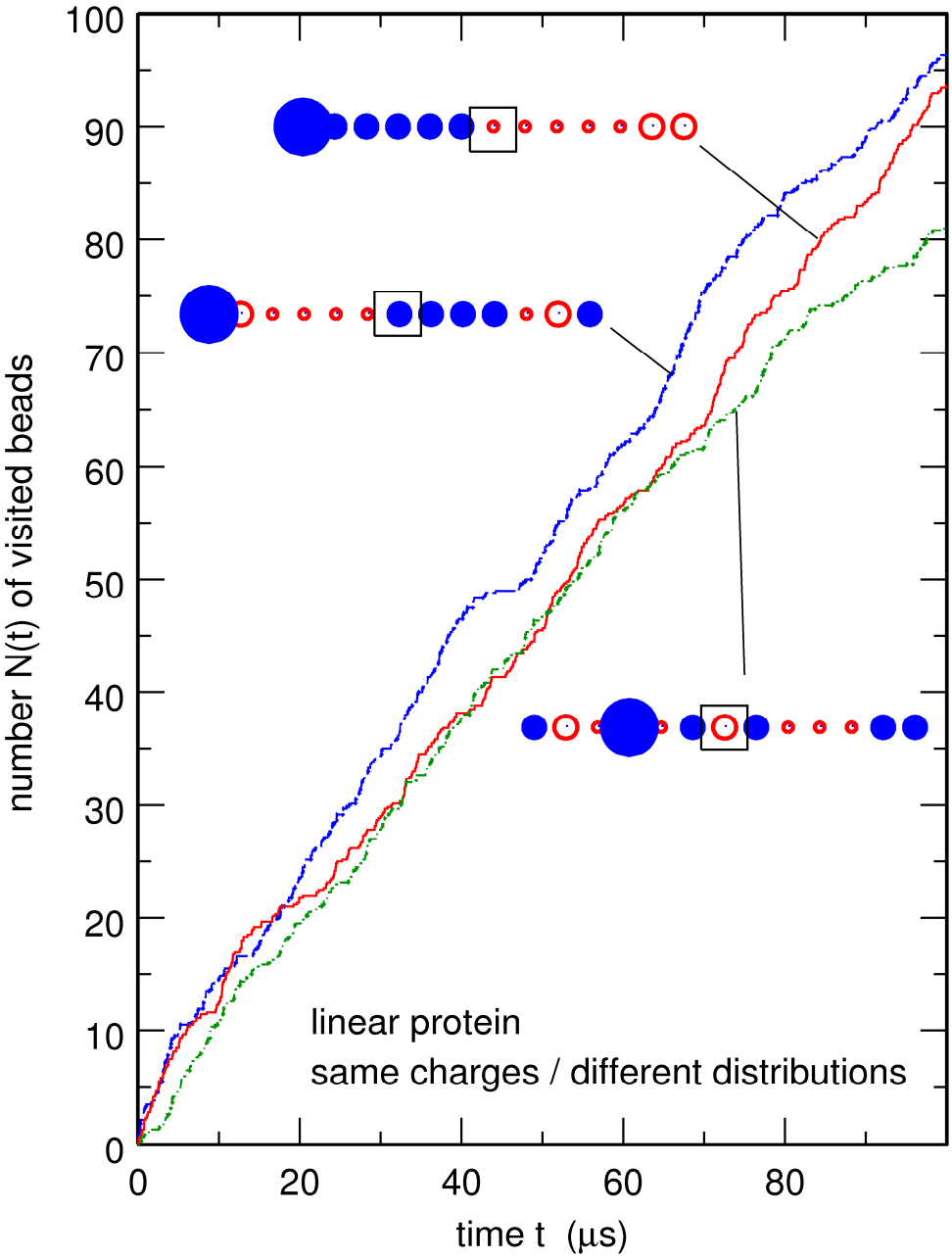

Figure 7.13. *Time evolution of the number* $N(t)$ *of different DNA beads visited by the protein search site for linear proteins with a gradient distribution of charges with total charge* $e_{\text{prot}} = -2.4e_{\text{DNA}}$ *and maximum charge* $e_{\max} = -1.2e_{\text{DNA}}$ *(solid line), as well as two distributions obtained by random permutations of these charges. In the small inserts are shown the positions of the charges at equilibrium. Filled circles correspond to positive charges and empty ones to negative charges, the radius of each circle being proportional to the absolute value of the charge. The search site, which is surrounded by a square, is the central (seventh) bead of each chain. It has charge* $0.13e_{\text{DNA}}$ *for the gradient distribution, and charges* $-0.53e_{\text{DNA}}$ *and* $0.40e_{\text{DNA}}$ *for the random distributions. It was considered that protein bead p is attached to bead k of DNA segment j if* $\left\| \mathbf{r}_{j,k} - \mathbf{R}_p \right\| \leq \sigma$.

8. Comparison of kinetic and dynamical models and discussion on facilitated diffusion

In the two previous chapters, I have computed the acceleration of DNA sampling due to facilitated diffusion. I have shown that for the single bead protein the maximum acceleration due to facilitated diffusion is not larger than two, a value that is much smaller than predicted by other models [11-19,97]. The results obtained with the second model for the protein confirm this hypothesis. Actually, for the 13 beads proteins there are several cases where the search speed is even below the diffusion limit. This means that there is no combination of 3D diffusion and 1D sliding that is faster than normal diffusion. This asks for a more detailed analysis of the concept of facilitated diffusion: whether it is really more efficient than normal diffusion and also what happens in real systems. Therefore, in this chapter I compare results obtained with dynamical and kinetic models, also taking into account some recent reviews of the results in this field, and then I present some conclusions regarding facilitated diffusion in real systems.

8.1. Methodology

To compare dynamical and kinetic models, I ran additional simulations for the single bead protein, where I varied the DNA concentration (through the variation of the parameter *w*, see chapter 6) and the protein charge. These simulations were done in two versions: including hydrodynamic interactions or ignoring them. As before, the most important quantity that I extract from the simulations is the total number of different DNA beads visited by the protein, *N(t)*. I checked that, for all of the investigated cases, the number of beads visited in time follows the law given by equation (6.20). By inverting this relation, one obtains that the mean time t_k of first arrival at the *k*th distinct bead is:

$$t_k = -\frac{mn}{\kappa}\ln\left(1-\frac{k}{mn}\right) \tag{8.1}$$

This relation is, however, necessarily wrong for the last DNA bead ($k = mn$), since it predicts that it takes an infinite time for the protein to reach this bead, while this time must be finite. By computing the mean time of first arrival τ over the other $mn-1$ beads, one obtains:

$$\tau = \frac{1}{mn-1}\sum_{k=1}^{mn-1} t_k = -\frac{1}{\kappa}\sum_{k=1}^{mn-1}\ln\left(1-\frac{k}{mn}\right) \tag{8.2}$$

which, for large values of mn, is very close to

$$\tau \approx \frac{mn}{\kappa} \tag{8.3}$$

It can be checked numerically that the validity of equation (8.3) degrades only slowly when the average in equation (8.2) is calculated over $mn-10$ or $mn-100$ beads instead of $mn-1$. This indicates that the validity of equation (8.3) does not depend too sensitively on the exact asymptotic behavior of $N(t)$ close to mn (remember that equation (6.20) remains valid even when the protein has already visited more than 99.5% of the total number of DNA beads).

Moreover, it is also possible to check that, for a pure diffusive 3D motion, the rate κ obtained from the time evolution of $N(t)$ and the mean time of first arrival τ obtained from Klenin *et al*'s formula in equation (5.9) are related through equation (8.3). Indeed, in the absence of sliding ($\tau_{1D} \to 0$) and for a radius *a* equal to δ, the mean time of first arrival obtained from Klenin *et al*'s relation in equation (5.9) tends towards:

$$\tau = \frac{V}{4\pi D_{3D}\delta} = \frac{mn}{4\pi D_{3D}\delta c} \tag{8.4}$$

When comparing equation (8.4) with equation (6.21), one finds the confirmation that κ and τ are related through equation (8.3) for 3D diffusion.

The strategy that I have adopted to compare my model with that of Klenin *et al* therefore consists in extracting several quantities from the simulations I ran. On one side, I have directly estimated the rate constant κ from each simulation by fitting the computed evolution of $N(t)$ against equation (6.20). On the other side, I have also derived numerical values for D_{1D}, D_{3D}, τ_{1D} and τ_{3D} from the same simulations (see below for more detail). I used these values to compute the mean time of first arrival τ according to Klenin *et al*'s formula in equation (5.9). Finally, I converted τ to a rate constant κ by using equation (8.3) and compared it to the value of κ deduced from the time evolution of $N(t)$.

8.2. Computation of the quantities needed to compare dynamical and kinetic models

Figure 8.1 displays a logarithmic plot of the time evolution of $1-N(t)/(mn)$, with $mn=4000$, the fraction of DNA beads not yet visited by the protein, for $e_{\text{prot}}/e_{\text{DNA}}=1$ and for values of *w* ranging between 18 nm and 135 nm (with hydrodynamic interactions). Rate constants κ were extracted from such plots by fitting the computed evolution of $N(t)$ against equation (6.20). These values are reported in table 8.1 in units of beads/μs. This table has 24 entries, which correspond to all possible combinations obtained with four values of *w* (18, 32, 45 and 135 nm), three different DNA-protein interaction laws (repulsive interactions, $e_{\text{prot}}/e_{\text{DNA}}=1$, and $e_{\text{prot}}/e_{\text{DNA}}=3$), and two different ways of handling hydrodynamic interactions ("off" and "on"). As will also be the case for all subsequent tables, the first number in each entry was obtained with the $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{\text{prot}}\right\|\le\sigma$ criterion, while the number in parentheses was obtained with the $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{\text{prot}}\right\|\le 1.5\sigma$ criterion. It is seen that the values of κ vary over more than two orders of magnitude and depend very strongly on whether hydrodynamic interactions are taken into account or not.

Klenin *et al*'s formula in equations (5.9) and (5.10) depends on $\tau_{1\text{D}}$ and $\tau_{3\text{D}}$, the average times the protein spends in the bound and free states, respectively. Equation (5.10) may be rewritten in the slightly more convenient form

$$\xi = w\sqrt{\frac{1}{2\pi}\frac{D_{1\text{D}}}{D_{3\text{D}}}\frac{\rho_{1\text{D}}}{1-\rho_{1\text{D}}}} \tag{8.5}$$

where $\rho_{1\text{D}}$ denotes the fraction of time during which the protein is attached to a DNA bead, that is $\rho_{1\text{D}}=\tau_{1\text{D}}/(\tau_{1\text{D}}+\tau_{3\text{D}})$. Values of $\rho_{1\text{D}}$ are easily extracted from the simulations by checking at each time step whether the distance between the center of the protein bead and that of any DNA bead is smaller than the threshold, that is σ or 1.5σ. The obtained values of $\rho_{1\text{D}}$ are shown in table 8.2. As already emphasized in the preceding chapters, $\rho_{1\text{D}}$ increases from nearly 0 for the

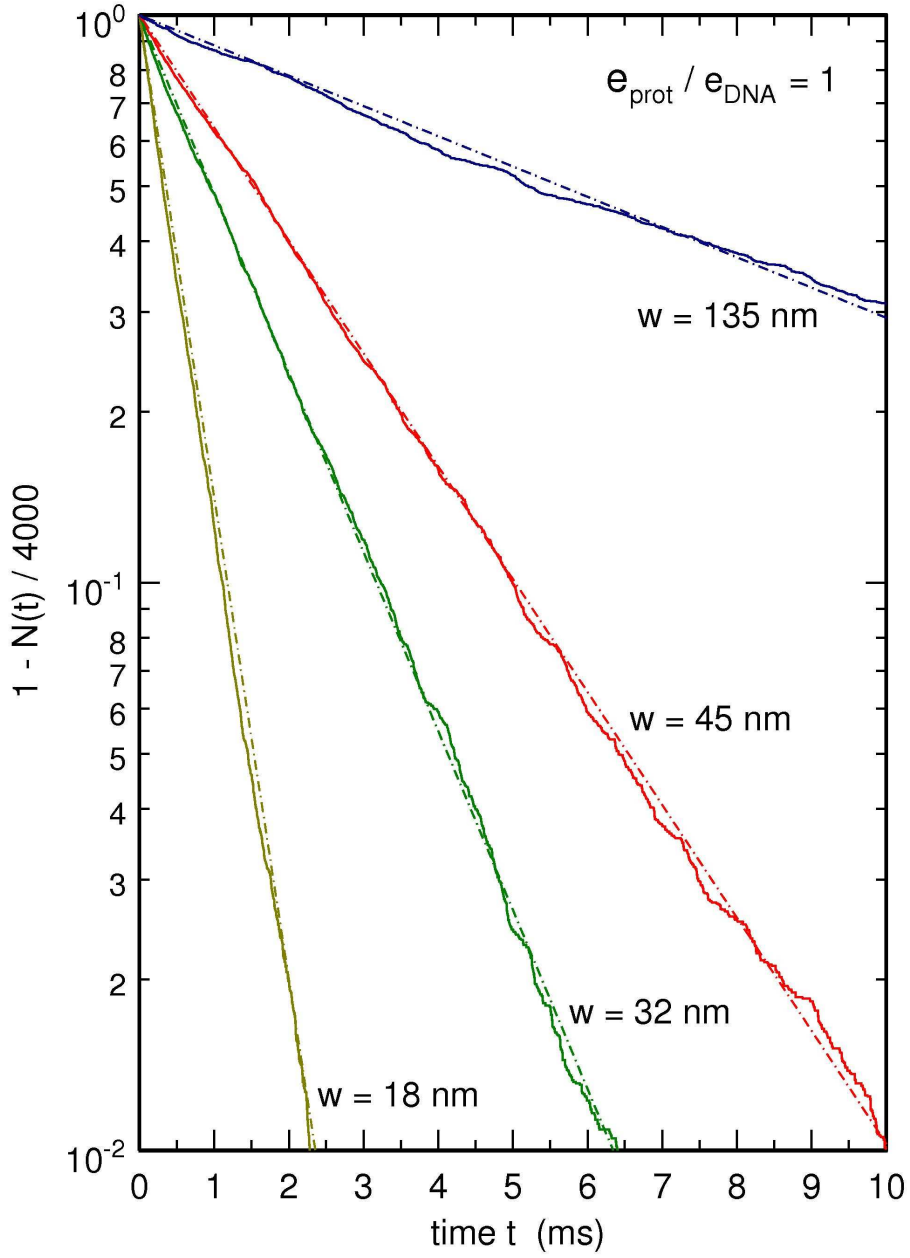

Figure 8.1. *Logarithmic plot of the time evolution of* $1-N(t)/4000$ *, the fraction of DNA beads not yet visited by the protein, for* $e_{\mathrm{prot}}/e_{\mathrm{DNA}}=1$ *and four values of w ranging between 18 nm and 135 nm. Hydrodynamic interactions are taken into account. It was furthermore considered that the protein is attached to bead k of DNA segment j if* $\left\|\mathbf{r}_{j,k}-\mathbf{r}_{\mathrm{prot}}\right\|\leq\sigma$ *. The dot-dashed straight lines, which were adjusted against the evolution of* $1-N(t)/4000$ *for each value of w, were used to estimate the values of κ.*

HI	w (nm)	κ (units of beads/μs)		
		repulsive potential	$e_{\text{prot}} / e_{\text{DNA}} = 1$	$e_{\text{prot}} / e_{\text{DNA}} = 3$
off	18	2.70 (3.86)	2.32 (2.69)	0.30 (0.34)
	32	0.98 (1.44)	0.84 (0.91)	0.121 (0.127)
	45	0.47 (0.70)	0.52 (0.55)	0.086 (0.089)
	135	0.050 (0.075)	0.149 (0.153)	0.037 (0.038)
on	18	5.73 (8.40)	7.82 (10.30)	7.76 (8.96)
	32	1.94 (3.00)	2.90 (3.43)	2.85 (3.10)
	45	1.08 (1.68)	1.83 (2.11)	1.59 (1.70)
	135	0.30 (0.38)	0.49 (0.53)	0.40 (0.41)

Table 8.1. *Values of the rate constant κ (expressed in units of beads/μs), obtained by fitting the time evolution of $N(t)$ against equation (6.20), for different values of w, different DNA-protein interaction laws, and hydrodynamic interactions switched either "off" or "on". The first number in each entry was obtained with the $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \le \sigma$ criterion, while the number in parentheses was obtained with the $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \le 1.5\sigma$ criterion.*

HI	w (nm)	$\rho_{1\text{D}}$		
		repulsive potential	$e_{\text{prot}} / e_{\text{DNA}} = 1$	$e_{\text{prot}} / e_{\text{DNA}} = 3$
off	18	0.12 (0.43)	0.60 (0.982)	0.912 (1.000)
	32	0.04 (0.16)	0.60 (0.961)	0.902 (1.000)
	45	0.02 (0.09)	0.61 (0.995)	0.906 (1.000)
	135	< 0.01 (0.01)	0.29 (0.44)	0.46 (0.56)
on	18	0.15 (0.44)	0.32 (0.74)	0.66 (0.985)
	32	0.05 (0.17)	0.23 (0.53)	0.66 (0.979)
	45	0.03 (0.11)	0.20 (0.41)	0.67 (0.986)
	135	< 0.01 (0.01)	0.09 (0.19)	0.56 (0.78)

Table 8.2. *Values of $\rho_{1\text{D}}$, the fraction of time during which the protein is attached to a DNA bead, for different values of w, different DNA-protein interaction laws, and hydrodynamic interactions switched either "off" or "on". The first number in each entry was obtained with the $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \le \sigma$ criterion, while the number in parentheses was obtained with the $\left\|\mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}}\right\| \le 1.5\sigma$ criterion.*

repulsive potential to almost 1 for large values of the protein charge. It can also be seen in table 8.2 that ρ_{1D} is substantially smaller when hydrodynamic interactions are taken into account than when they are not. Stated in other words, hydrodynamic interactions (HI) tend to move the protein away from the DNA. We will see that this has a marked effect on the targeting speed. At last, it can also be noticed that the values of ρ_{1D} for the largest value of *w* (145 nm) are substantially smaller than for the three other values of *w* (18, 32, and 45 nm), which reflects the fact that DNA segments are more widely separated and the protein consequently spends more time diffusing freely in the buffer.

In table 8.3 are given the values of the 1D diffusion coefficients D_{1D} in units of $10^{-10}m^2s^{-1}$. They were computed, as in the previous chapter, by drawing log-log plots of the average value of the number of visited beads during long sliding events. A few representative plots are shown in figure 8.2. All plots are approximately linear in log-log scales, which means that $N(t)$ evolves according to a power law $N(t) \approx \alpha t^{\beta}$. I found that β is close to 0.5 for $e_{prot}/e_{DNA} = 1$ and HI switched "on", to 0.45 for $e_{prot}/e_{DNA} = 1$ and HI switched "off", to 0.40 for $e_{prot}/e_{DNA} = 3$ and HI switched "on", and to 0.20 for $e_{prot}/e_{DNA} = 3$ and HI switched "off". This indicates that the sliding motion is diffusive in the first case, slightly subdiffusive in the second and third cases, and very subdiffusive in the last case. This is probably connected to the fact that, when going from the first to the fourth case, the protein bead actually spends more and more time attached to the same DNA bead without moving: large average waiting times between random-walk steps are indeed sufficient to induce subdiffusion (see, for example, [145]). Except for the last case, the time evolution of $N(t)$ can therefore be fitted with a square-root law $N(t) \approx \alpha\sqrt{t}$ and then the diffusion coefficient is obtained using equation (5.12) and the relation $\ell(t) = l_0 N(t)$. As could reasonably be expected, the estimated values of D_{1D} do not depend on the value of *w*. In contrast, D_{1D} appears to be about twice larger when HI are taken into account than when they are not. Not surprisingly, D_{1D} also depends to some extent on the shape and depth of the interaction potential: values of D_{1D} for $e_{prot}/e_{DNA} = 3$ appear to be about 40% larger than the corresponding values for $e_{prot}/e_{DNA} = 1$.

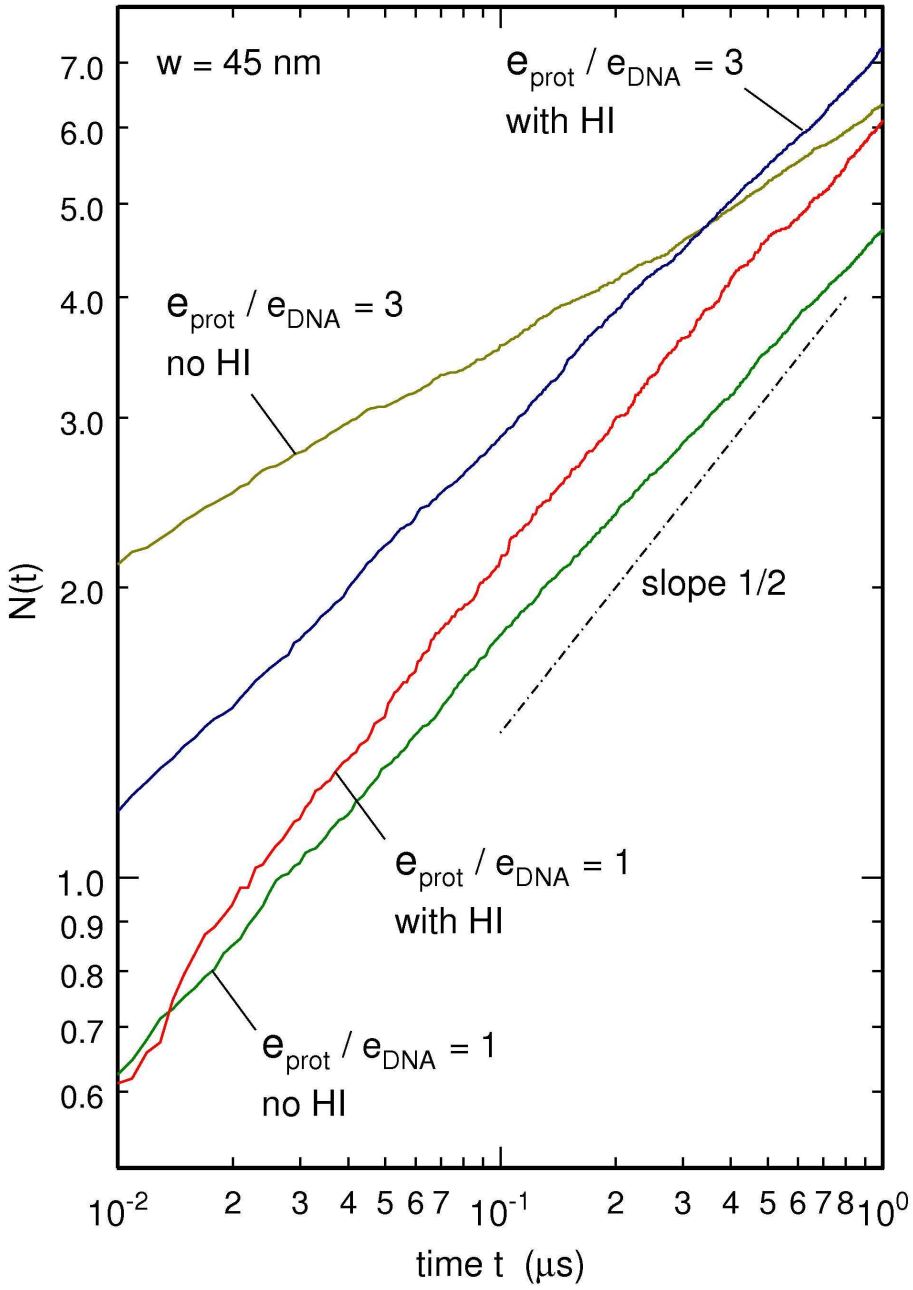

Figure 8.2. *Log-log plots of the time evolution of the number* $N(t)$ *of different DNA beads visited by the protein during 1D sliding for various systems with w=45 nm. As indicated on the figure, two simulations were ran with* $e_{\text{prot}} / e_{\text{DNA}} = 1$ *and two other ones with* $e_{\text{prot}} / e_{\text{DNA}} = 3$. *Similarly, hydrodynamic interactions were taken into account for two of the simulations, but neglected for the two other ones. It was considered that the protein is attached to bead k of DNA segment j if* $\left\| \mathbf{r}_{j,k} - \mathbf{r}_{\text{prot}} \right\| \leq \sigma$. *For each simulation,* $N(t)$ *was averaged over several tens of sliding events with the following properties : (i) each sliding event lasted more than 1 μs, (ii) the protein did not separate from the DNA segment by more than* σ *during more than 0.07 μs, (iii) the protein bead did not reach one of the extremities of the DNA segment.*

HI	*w* (nm)	D_{1D} (units of 10^{-10} m^2 s^{-1})	
		$e_{prot}/e_{DNA}=1$	$e_{prot}/e_{DNA}=3$
off	18	1.15 (1.30)	
	32	1.18 (1.21)	
	45	1.14 (1.20)	
	135	1.15 (1.21)	
on	18	1.94 (2.29)	3.13 (3.18)
	32	2.15 (2.71)	2.82 (2.54)
	45	1.93 (2.74)	2.72 (2.62)
	135	1.92 (2.39)	2.45 (2.11)

Table 8.3. *Values of D_{1D}, the diffusion coefficient of the protein along the DNA segment, expressed in units of 10^{-10} m^2 s^{-1}, for different values of w, different DNA-protein interaction laws, and hydrodynamic interactions switched either "off" or "on". The first number in each entry was obtained with the $\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\|\le\sigma$ criterion, while the number in parentheses was obtained with the $\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\|\le 1.5\sigma$ criterion. 1D motion for $e_{prot}/e_{DNA}=3$ and HI switched "off" is too subdiffusive to be described by a diffusion coefficient D_{1D}.*

HI	*w* (nm)	D_{3D} (units of 10^{-10} m^2 s^{-1})
off	18	0.66 (0.63)
	32	0.73 (0.72)
	45	0.72 (0.71)
	135	0.69 (0.69)
on	18	1.40 (1.37)
	32	1.45 (1.49)
	45	1.65 (1.71)
	135	4.11 (3.47)

Table 8.4. *Values of D_{3D}, the diffusion coefficient of the protein in the buffer, expressed in units of 10^{-10} m^2 s^{-1}, for different values of w, and hydrodynamic interactions switched either "off" or "on". The first number in each entry was obtained with the $\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\|\le\sigma$ criterion, while the number in parentheses was obtained with the $\|\mathbf{r}_{j,k}-\mathbf{r}_{prot}\|\le 1.5\sigma$ criterion. The values of D_{3D} were obtained from the expression of the volume of the 3D Wiener sausage in equation (6.21) and the values of κ reported in the "repulsive potential" column of table 8.1.*

As already shown before, the 3D diffusion coefficient of the protein in the buffer can be estimated in at least three different ways, namely from Einstein's formula (equation (6.23) – in the case of a single bead protein this gives $D_{3D} = 0.70\times10^{-10}$ m^2 s^{-1}), from the expression of the volume of the 3D Wiener sausage (equation (6.21)), and from the mean squared displacement of the protein (equation (5.11)). For repulsive DNA/protein interactions and HI switched "off", the values of D_{3D} obtained with these methods should be very close. The estimates obtained from the values of κ in the "repulsive potential" column of table 8.1 and equation (6.21) with $\delta = \sigma$ or $\delta = 1.5\sigma$ are shown in the top half of table 8.4. It can be checked that they indeed agree very closely with the result of Einstein's formula. In contrast, when HI are taken into account, the values of D_{3D} obtained from the expression of the volume of the 3D Wiener sausage are substantially larger than those obtained from Einstein's formula (see the bottom half of table 8.4). This agrees with Kirkwood-Riseman's equation, which states that HI reduce the effective friction coefficient of long DNA chains [148]. However, this is in apparent contradiction with other works that state that HI tend to decrease the association rate between two diffusing spheres placed at short distance [149-151], because the stochastic (thermal) motions of the two particles become highly correlated, which slows down their relative mobility. I therefore confirmed this result by extracting D_{3D} from the time evolution of the mean squared displacement of the protein, according to equation (5.11), that is, more precisely:

$$\left\langle \left\| \mathbf{r}_{\text{prot}}(t) - \mathbf{r}_{\text{prot}}(0) \right\|^2 \right\rangle = 6D_{3D}t\,, \tag{8.6}$$

For example, I checked that equation (8.6) leads to $D_{3D} = 0.68\times10^{-10}$ m^2 s^{-1} for w=45 nm and HI switched "off" and to $D_{3D} = 1.60\times10^{-10}$ m^2 s^{-1} for HI switched "on". The dependence of the 3D diffusion coefficient of the protein on HI is a point that certainly deserves further attention on its own.

All the quantities that are necessary to estimate the rate constant κ from Klenin *et al*'s formula for the mean time of first arrival τ in equations (5.9) and (8.5) and the relation between τ and κ in equation (8.3) are now at disposal. These values of κ are reported in table 8.5 in units of beads/μs. Since there is no sliding of the protein along the DNA for the repulsive DNA/protein interaction, ρ_{1D} was set to 0 in this case in Klenin *et al*'s formula, although ρ_{1D} is actually small

but not zero because of collisions (see table 8.2). As a consequence, the "repulsive potential" column of table 8.5 is similar to that of table 8.1, because this column of table 8.1 is used to estimate the 3D diffusion coefficient D_{3D} (table 8.4) according to the expression for the volume of the 3D Wiener sausage in equation (6.21). Moreover, for $e_{prot}/e_{DNA}=3$ and HI switched "off", the sliding motion of the protein along DNA is too subdiffusive to enable an estimation of D_{1D}. Klenin *et al*'s formula can therefore not be used in this latter case.

8.3. Acceleration of targeting due to facilitated diffusion and hydrodynamic interactions

Table 8.6 shows the acceleration of the protein targeting process due to facilitated diffusion. This acceleration was estimated as the ratio of a given rate constant κ for $e_{prot}/e_{DNA}=1$ or $e_{prot}/e_{DNA}=3$ divided by the corresponding value of κ for the repulsive DNA/protein interaction. Table 8.7 similarly shows the acceleration of the targeting process due to HI. This acceleration was estimated as the ratio of a given rate constant κ for HI switched "on" divided by the corresponding value of κ for HI switched "off". In both cases, the values of κ were taken from table 8.1 for the dynamical model and from table 8.5 for the kinetic model.

Let us first concentrate on the results obtained with the dynamical model. For a reason which will become clear later, I first discuss the results obtained with the σ threshold. For HI switched "on", the values for the acceleration of targeting due to facilitated diffusion reported in table 8.6 are comprised between 1.3 and 1.7 and are quite similar for $e_{prot}/e_{DNA}=1$ and $e_{prot}/e_{DNA}=3$ (remember that the acceleration becomes smaller than 1 for values of e_{prot}/e_{DNA} larger than 5). Table 8.6 additionally indicates that the acceleration due to facilitated diffusion depends only marginally on *w*, and consequently on DNA concentration, when HI are considered. Things are, however, quite different when HI are switched "off". In this case, the acceleration due to facilitated diffusion depends significantly on *w*. When *w* increases from 18 nm to 135 nm, the acceleration indeed increases by a factor of almost 4 for $e_{prot}/e_{DNA}=1$, and almost 7 for $e_{prot}/e_{DNA}=3$. Moreover, the value of the acceleration depends much more sharply on the

HI	*w* (nm)	κ (units of beads/µs)		
		repulsive potential	$e_{\text{prot}} / e_{\text{DNA}} = 1$	$e_{\text{prot}} / e_{\text{DNA}} = 3$
off	18	2.70 (3.86)	2.16 (0.48)	
	32	0.98 (1.44)	1.08 (0.38)	
	45	0.47 (0.70)	0.70 (0.09)	
	135	0.050 (0.075)	0.21 (0.23)	
on	18	5.73 (8.40)	5.15 (3.93)	4.43 (0.95)
	32	1.94 (3.00)	2.35 (2.76)	2.25 (0.59)
	45	1.08 (1.68)	1.48 (2.02)	1.45 (0.36)
	135	0.30 (0.38)	0.52 (0.69)	0.80 (0.57)

Table 8.5*. Values of the rate constant κ (expressed in units of beads/µs) obtained from Klenin et al's formula for τ in equations (5.9) and (8.5), the relation between τ and κ in equation (8.3), and the values of* ρ_{1D}, D_{1D} *and* D_{3D} *in tables 8.2 to 8.4, for different values of w, different DNA-protein interaction laws, and hydrodynamic interactions switched either "off" or "on". Since there is no sliding of the protein along the DNA for the repulsive DNA/protein interaction,* ρ_{1D} *was set to 0 in this case in Klenin et al's formula, although* ρ_{1D} *is actually small but not zero (see table 8.2). Moreover, for* $e_{\text{prot}} / e_{\text{DNA}} = 3$ *and hydrodynamic interactions switched "off", the sliding motion of the protein along the DNA is too subdiffusive to enable an estimation of* D_{1D} *. Klenin et al's formula can therefore not be used in this latter case.*

HI	*w* (nm)	$e_{\text{prot}} / e_{\text{DNA}} = 1$		$e_{\text{prot}} / e_{\text{DNA}} = 3$	
		dynamical	kinetic	dynamical	kinetic
off	18	0.86 (0.70)	0.80 (0.12)	0.11 (0.09)	
	32	0.86 (0.63)	1.10 (0.26)	0.12 (0.09)	
	45	1.10 (0.79)	1.49 (0.13)	0.18 (0.13)	
	135	2.98 (2.04)	4.20 (3.07)	0.74 (0.51)	
on	18	1.36 (1.23)	0.90 (0.47)	1.35 (1.07)	0.77 (0.11)
	32	1.49 (1.14)	1.21 (0.92)	1.47 (1.03)	1.16 (0.20)
	45	1.69 (1.26)	1.37 (1.20)	1.47 (1.01)	1.34 (0.21)
	135	1.63 (1.39)	1.73 (1.82)	1.33 (1.08)	2.67 (1.50)

Table 8.6. *Acceleration of the protein targeting process due to facilitated diffusion, for both the dynamical and kinetic models, estimated as the ratio of a given rate constant κ for* $e_{\text{prot}} / e_{\text{DNA}} = 1$ *or* $e_{\text{prot}} / e_{\text{DNA}} = 3$ *divided by the corresponding value of κ for the repulsive DNA/protein interaction. The values of κ were taken from table 8.1 for the dynamical model and from table 8.5 for the kinetic model.*

w (nm)	repulsive potential		$e_{\text{prot}}/e_{\text{DNA}}=1$		$e_{\text{prot}}/e_{\text{DNA}}=3$	
	dynamical	kinetic	dynamical	kinetic	dynamical	kinetic
18	2.12 (2.18)	2.12 (2.18)	3.37 (3.83)	2.38 (8.19)	25.9 (26.4)	
32	1.98 (2.08)	1.98 (2.08)	3.45 (3.76)	2.18 (7.26)	23.6 (24.4)	
45	2.30 (2.40)	2.30 (2.40)	3.52 (3.84)	2.11 (22.4)	18.5 (19.1)	
135	6.00 (5.07)	6.00 (5.07)	3.29 (3.46)	2.48 (2.26)	10.8 (10.8)	

Table 8.7. *Acceleration of the protein targeting process due to hydrodynamic interactions (HI), for both the dynamical and kinetic models. Acceleration of targeting due to HI was estimated as the ratio of a given rate constant κ for HI switched "on" divided by the corresponding value of κ for HI switched "off". In both cases, the values of κ were taken from table 8.1 for the dynamical model and from table 8.5 for the kinetic model.*

protein charge than for HI switched "on". Indeed, in the range of values of w I investigated, acceleration of targeting for $e_{\text{prot}}/e_{\text{DNA}}=1$ is larger than that for $e_{\text{prot}}/e_{\text{DNA}}=3$ by a factor which varies between 3.5 and 8. More precisely, facilitated diffusion is about 10 times *slower* than 3D diffusion for $e_{\text{prot}}/e_{\text{DNA}}=3$ and w=18 nm, but more than 3 times *faster* for $e_{\text{prot}}/e_{\text{DNA}}=1$ and w=135 nm.

The crucial role of hydrodynamics is further emphasized by the values of the acceleration of targeting due to HI reported in table 8.7. It is seen that, for values of w close to physiological ones (30 to 50 nm), this acceleration is close to 2 for repulsive DNA/protein interactions and to 3.5 for $e_{\text{prot}}/e_{\text{DNA}}=1$, while it is as large as 20 for $e_{\text{prot}}/e_{\text{DNA}}=3$. Examination of tables 8.2 to 8.4 suggests that the large acceleration of targeting observed when HI are switched "on" is ascribable to two rather distinct effects. First, as already noted in the preceding section, both $D_{1\text{D}}$ and $D_{3\text{D}}$ are roughly twice larger when HI are switched "on" than when they are switched "off" (see tables 8.3 and 8.4). This, of course, accelerates the targeting process in proportion. The second effect is that HI tend to detach the protein from the DNA sequence, as can be checked by looking at the values of $\rho_{1\text{D}}$ reported in table 8.2. This considerably modifies the motion of highly charged proteins. For example, for $e_{\text{prot}}/e_{\text{DNA}}=3$ and HI switched "off", the protein spends about 90% of the time attached to DNA for physiological values of w. The protein remains consequently attached for most of the time to the same portion of the DNA sequence and

either does not move or performs essentially 1D search, which is quite inefficient (see equation (5.12)). In contrast, ρ_{1D} is of the order of 66% when HI are switched "on", so that, in spite of the strong electrostatic attraction exerted by DNA, the protein spends a sizeable amount of time diffusing in 3D in the buffer. Stated in other words, the reduction of ρ_{1D} caused by HI allows strongly charged proteins to search efficiently for their target, while this would be forbidden by electrostatic interactions in the absence of HI.

8.4. Comparison of the dynamical and kinetic models

Let us now examine the agreement between results obtained with the dynamical and kinetic models, and let us start with the results obtained when switching HI "off". For the repulsive DNA/protein interaction potential, the corresponding columns of tables 8.1 and 8.5 are identical. This actually just reflects the facts that the values of κ in table 8.1 were used to estimate the diffusion coefficients D_{3D} reported in table 8.4 and that ρ_{1D} was further assumed to be zero in equation (8.5) for repulsive DNA/protein interactions, because in this case it is not possible to derive an estimation of D_{1D} from Brownian dynamics simulations. Still, when plugging in equation (8.4) the value of D_{3D} obtained from Einstein formula (equation (6.23)) instead of those reported in table 8.4, one again obtains "kinetic" rate constants κ that are in excellent agreement with "dynamical" ones.

While for repulsive DNA/protein interactions the agreement between the dynamical and kinetic models does not depend on the threshold used in Brownian dynamics simulations, this is no longer the case for the interaction potential with $e_{prot}/e_{DNA}=1$. Comparison of tables 8.1 and 8.5 indeed indicates that the agreement is pretty good for the σ threshold, while the values of κ estimated from Klenin *et al*'s formula are much too small for the 1.5σ threshold. This is actually also the case for all the simulations that will be discussed in the remainder of this chapter. Examination of tables 8.3 and 8.4 indicates that the values of D_{1D} and D_{3D} derived from Brownian dynamics simulations are not sensitive to the threshold, as one would reasonably expect. In contrast, the fraction of time ρ_{1D} during which the protein is attached to the DNA sequence depends strongly on the threshold. In particular, the 1.5σ threshold leads to values of

ρ_{1D} that are close to 1 for most of the simulations. The point is, that the values of the rate constant κ obtained from Klenin *et al*'s formula tend towards 0 when ρ_{1D} tends towards 1. This reflects the fact that the protein motion thereby switches from facilitated diffusion, for which $N(t)$ increases linearly with time, to 1D diffusion, for which $N(t)$ increases as the square root of time. Overestimation of ρ_{1D} therefore essentially results in underestimation of κ. This is very clearly what happens when the 1.5σ threshold is used in Brownian dynamics simulations. In contrast, it seems that the σ threshold leads to values of ρ_{1D} that perform a better job as input values to Klenin *et al*'s formula. Therefore, I will henceforth only consider values obtained with the σ threshold.

The top half of the last column of table 8.5 is void. This is due to the fact that the sliding motion of the protein for $e_{prot}/e_{DNA} = 3$ and HI switched "off" is so much subdiffusive that it is neither meaningful nor practically feasible to extract diffusion coefficients D_{1D} from the simulations. As a direct consequence, it is not possible in this case to derive estimates of κ from Klenin *et al*'s formula. I am not familiar enough with the theoretical background of reference [117] to determine whether this is a fundamental limitation of the kinetic model, or whether equations (5.9) and (5.10) can be generalized to account for subdiffusive 1D motion of the protein.

Let us now compare results obtained with the dynamical and kinetic models when HI are switched "on". The kinetic model does not explicitly incorporate them, which rises an interesting question: are HI reducible to their effect on D_{1D}, D_{3D}, and ρ_{1D}? Stated in other words, is it sufficient to plug in Klenin *et al*'s expression the values of D_{1D}, D_{3D}, and ρ_{1D} deduced from simulations with HI switched "on" to get reasonable estimates of κ? Comparison of the bottom halves of tables 8.1 and 8.5 suggests that this is indeed the case. Even if the values of κ differ in one case by a factor of 2, the agreement is generally correct.

8.5. What about real systems?

In this chapter, I have thus shown that the dynamical model I proposed and the kinetic model of Klenin *et al* [118] support each other, in the sense that the rate constants κ obtained (i) directly from the simulations, and (ii) from Klenin *et al*'s formula using values of D_{1D}, D_{3D}, and ρ_{1D} extracted from the simulations, are in good agreement. In particular, both models suggest that the acceleration of targeting due to facilitated diffusion is not very large for the system I considered. Table 8.6 indeed shows that the dynamical and kinetic models agree in predicting an acceleration comprised between 20% and 70% for physiological values of *w*, HI switched "on", and protein charges ranging from $e_{prot}/e_{DNA} = 1$ to $e_{prot}/e_{DNA} = 3$.

However, one must at this point wonder how this result transfers to real DNA and proteins. The essential point is that the dynamical system corresponds to a ratio D_{1D}/D_{3D} of the order of unity (see tables 8.3 and 8.4), as is customary for translational diffusion. In contrast, the ratio D_{1D}/D_{3D} for real DNA/protein systems (measured essentially by single molecule experiments) is rather of the order of $\approx 10^{-3}$ [99,100,102,103,105,106,152-154]. This three orders of magnitude difference may be due to the fact that in real systems the protein has to follow a helical track along the DNA, which considerably enhances the translational friction coefficient [98,101,155]. Using Klenin *et al*'s formula, acceleration of targeting due to facilitated diffusion can be written in the form:

$$\frac{\kappa}{4\pi D_{3D} a c} = \frac{1-\rho_{1D}}{\frac{\pi}{2}\frac{a}{\xi}[1-\frac{2}{\pi}\arctan(\frac{a}{\xi})]} \tag{8.7}$$

where ξ is given in equation (8.5) and *a* is taken here as the sum of the protein and DNA hydrodynamic radii, σ. As a consequence, for a given DNA concentration (and therefore a given value of *w*), the acceleration due to facilitated diffusion depends uniquely on ρ_{1D} and the ratio D_{1D}/D_{3D}. For each value of D_{1D}/D_{3D}, one can therefore search for the value of ρ_{1D} for which this acceleration is maximum. The result is plotted in figure 8.3 for three different values of *w* (18, 45 and 135 nm). The top plots show the largest acceleration of targeting (relative to 3D diffusion) that can be attained for each value of D_{1D}/D_{3D}, and the bottom plots the value of ρ_{1D}

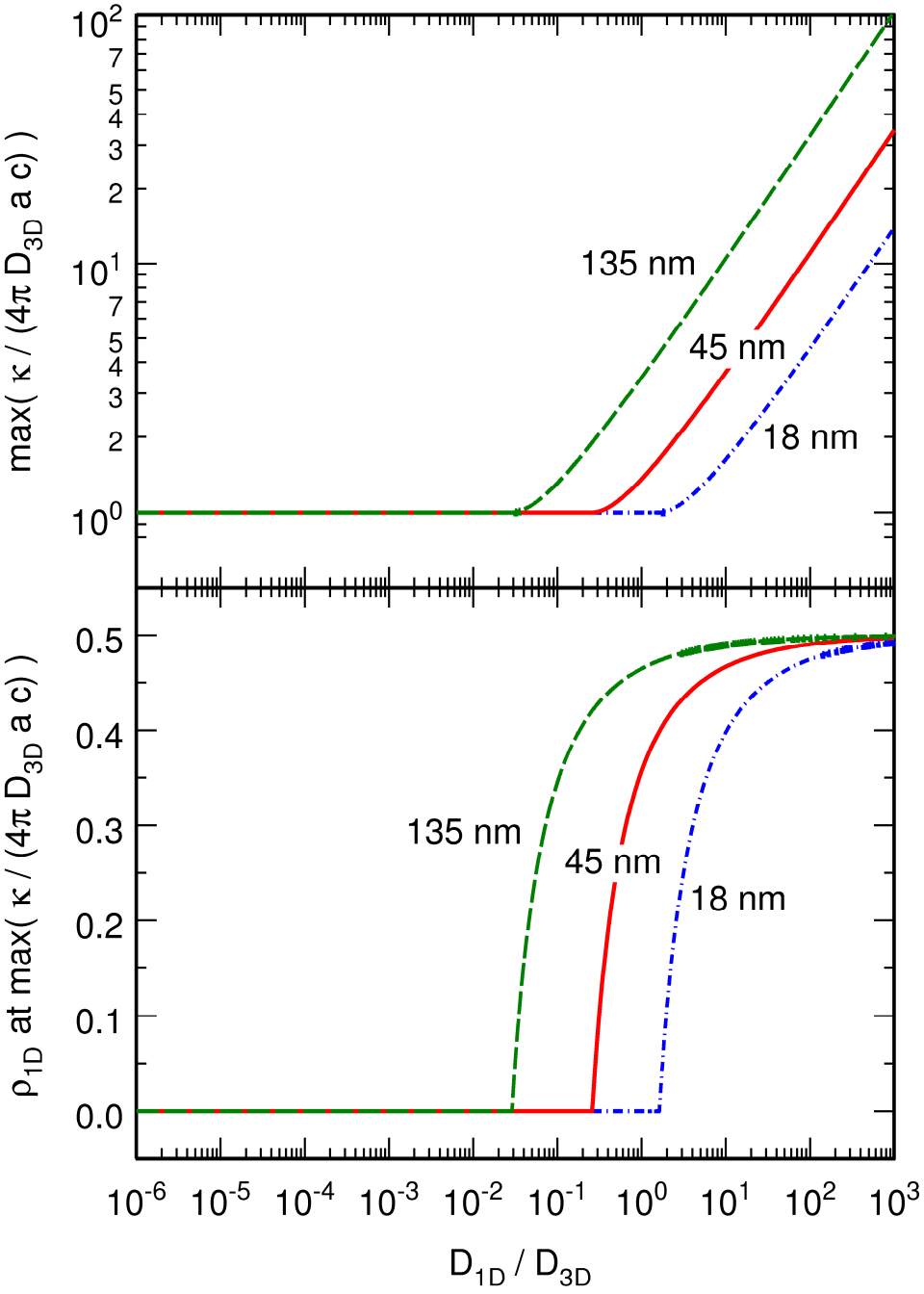

Figure 8.3. *Plot, as a function of* D_{1D} / D_{3D} *and for three different values of w (18, 45 and 135 nm), of the maximum value of* $\kappa/(4\pi D_{3D} a\, c)$ *that can be attained for values of* ρ_{1D} *comprised between 0 and 1 (top plot), and plot of the value of* ρ_{1D} *at which this maximum is attained (bottom plot).* $\kappa/(4\pi D_{3D} a\, c)$ *is evaluated according to equation (8.7). This ratio represents the maximum value of the acceleration of targeting, compared to 3D diffusion, which can be achieved thank to facilitated diffusion.*

at which this maximum is attained. It is seen that, for physiological values of *w* (30-50 nm), facilitated diffusion cannot be faster than 3D diffusion for values of D_{1D} / D_{3D} smaller than about 0.3: maximum acceleration is indeed 1 at $\rho_{1D} = 0$. For values of D_{1D} / D_{3D} larger than this threshold, the maximum acceleration instead increases approximately as the square root of D_{1D} / D_{3D}. This maximum acceleration is furthermore attained for values of ρ_{1D} close to 1/2 when D_{1D} / D_{3D} is larger than about 1. At last, the maximum acceleration due to facilitated diffusion increases slowly with *w*.

For values of D_{1D} / D_{3D} close to 1.5, as in these simulations (see tables 8.3 and 8.4), figure 8.3 indicates that maximum acceleration due to facilitated diffusion is of the order of 2 for physiological values of *w*, which is exactly what I obtained (see table 8.6). In contrast, realistic values of D_{1D} / D_{3D} are much smaller than the 0.3 threshold, which implies, as already stated, that facilitated diffusion is necessarily slower than 3D diffusion. For such small values of D_{1D} / D_{3D}, equation (8.7) actually reduces to:

$$\frac{\kappa}{4\pi D_{3D} a c} \approx 1 - \rho_{1D} \tag{8.8}$$

This conclusion agrees with experimental results, which indicate that the measured apparent diffusion coefficient of molecules that do not interact with chromatin or nuclear structures (like the green fluorescent protein or dextrans) range between 10^{-11} and 10^{-10} m^2 s^{-1} [152-154], depending on their size, as predicted by Einstein's formula, while that of biologically active molecules is instead usually reduced by a factor of 10-100 compared to this formula [155-159].

8.6. Conclusion

In this chapter, I have shown that my model and the kinetic model of Klenin *et al* agree in predicting that facilitated diffusion cannot be much faster than normal diffusion. I have computed the rate constant κ firstly directly from simulations and then from Klenin *et al*'s formula, by using values of D_{1D}, D_{3D} and ρ_{1D} computed with the dynamical model. For physiological DNA concentrations and realistic protein charges, the acceleration of targeting due to facilitated diffusion is in both cases smaller than 70%. Actually, the alternation of sliding and 3D diffusion

in the buffer can be faster than normal diffusion only for values of D_{1D}/ D_{3D}, which are much larger than those measured experimentally. These results come as a confirmation of Halford's analysis of experimental results dealing with protein-DNA non-specific interactions and of his conclusion: during the past 40 years, there may indeed have been some mistakes in the understanding of protein–DNA association kinetics.

9. Conclusions and perspectives

The purpose of my thesis work was the study of DNA models at different resolutions and of DNA-protein interactions and facilitated diffusion.

I started with the study of the simplest DNA models, namely statistical ones, and presented their application to 2D electrophoresis display. Using a code based on the open source program MeltSim, I showed that the results of genome separation experiments can be predicted with an accuracy that is higher than that of the measured values. I also pointed out that the use of simple expressions for the mobility of the sequences in the gels is sufficient to reach such an accuracy. Actually, my results also prove that nowadays the limiting step in separation problems is the reproducibility of the experimental procedure and not the validity of the model. Finally, I showed that the results of 2D display experiments are not sufficient to determine the best set of parameters for the modeling of fragments separation in the second dimension, and that additional detailed measurements of the mobility of a few sequences are necessary to achieve this goal.

I next studied DNA melting using a dynamical model. More precisely, I improved the set of parameters of the dynamical model developed in our group, in order to get a better agreement with experimental results, which were not taken into account until now, like the critical force needed to keep two DNA strands separated and the dependence of the critical temperature on the length of the sequence. This model has some similarity with statistical models, in the sense that it is based on site-dependent, finite stacking and pairing enthalpies. However, in contrast to statistical models, no explicit temperature dependence is plugged in the dynamical model. Instead of site dependent stacking entropies, temperature evolution is indeed governed by the shape of the stacking and pairing interactions. I compared the results obtained with the improved model with those of statistical models and found satisfactory agreement. I also studied the critical behavior of the new model and observed that, if one relies on the temperature evolution of the specific heat, then DNA denaturation looks like a first order phase transition in a rather broad temperature interval. Very close to the critical temperature, one however observes a crossover to a smoother regime (second order transition?). If one instead relies on the temperature evolution of the singular part of free energy of the system, then the order of the DNA melting transition depends on the anharmonicity of the stacking interaction: it is second order for an almost harmonic stacking potential, but looks first order for large anharmonicities. This is somewhat reminiscent of statistical models, which describe DNA denaturation as a phase transition, which order depends on the way the partition function of a loop and the loop closure exponent are

computed.

In the second part of my thesis, I proposed a dynamical model for the description of protein-DNA interaction and facilitated diffusion, which is based on a DNA model inspired from polymer physics. This model suggests that, although DNA sampling is performed via a succession of 3D motion in the buffer, 1D sliding along the DNA chain, short hops between neighboring sites, and intersegmental transfers, the global motion of the protein still looks diffusive-like. I computed the rate at which the protein scans the DNA sequence and studied how it is affected by the electrostatic and mechanical properties of the protein, like its charge distribution, its total charge, its shape, and its elasticity. I showed that the model predicts that facilitated diffusion accelerates sampling in a certain range of values of the charge of the protein. Moreover, for reasonable values of the total charge of the protein, the number of base pairs visited during a single sliding event is comparable to the values deduced from single molecule experiments, that is from a few tens to a few hundreds base pairs. I also studied the effect that hydrodynamic interactions have on the sampling process and showed that they can significantly increase the scanning rate. Finally, I compared the results obtained with the dynamical model with those obtained with the kinetic model of Klenin *et al*, and showed that both models agree in predicting that facilitated diffusion cannot push the speed of DNA sampling far beyond the diffusion limit. For realistic values of the 1D and 3D diffusion coefficients, facilitated diffusion is even most probably slower than normal diffusion. This result comes as an argument in the debate whether protein-DNA association is faster than diffusion, and supports a recent review of experimental work, which concludes that there is no known example of a protein that finds its target faster than diffusion and that we should put "an end to 40 years of mistakes in protein-DNA association kinetics" [21].

Even tough most of the results presented here are quite reliable, the model I proposed has several limitations. First, it predicts a 1D diffusion coefficient which is much larger than experimentally measured values. This is probably due to the fact that my model takes into account neither the helical structure of DNA nor the specific (sequence dependent) interactions between DNA and the protein, which slow down the sliding process. Moreover, hydrodynamic interactions are treated in a simplified way, disregarding the fluidity of hydration layers and short-range lubrication effects.

To my mind, the most important improvement one should bring to this model deals with

the resolution at which DNA is described. Ideally, an improved model should combine the properties of the dynamical models discussed in the two parts of this work. More precisely, it should be complex enough to describe both DNA at the scale of a single base pair and the diffusion of the double strand in the buffer. In contrast, the DNA models discussed in the first part of this work do not permit the study of the global motion of the sequence in the buffer and of its diffusion coefficient, while the model for DNA-protein interactions proposed in the second part is not detailed enough to describe individual base pairs and their opening as a consequence of either protein pulling or temperature increase. A model, which in my opinion could be applied to both DNA melting and the study of DNA-protein interactions, was proposed recently [160,161]. In this model, a sugar, a phosphate and a base are each described by one bead, so that six beads are needed for each base pair. The Hamiltonian of this model is rather similar to that used in the second part of my thesis. It includes stretching, bending and electrostatic interactions between the different beads. It also includes torsion and stacking interactions, which give DNA its helical structure with the major and minor grooves. Base pairing interactions are described by Lennard-Jones-type potentials. Moreover, the parameters of the model were fitted to reproduce both the correct denaturation curves and the persistence length of double-stranded DNA.

This improved description of DNA should of course be complemented by a better description of the protein. One should indeed switch from working with ‘generic’ proteins to specific structures. An interesting candidate could be a recent coarse-grained model, which uses one bead to describe each amino acid that composes a protein [162]. The initial configuration is built by placing the centers of the beads at the positions of the C_α atoms of the X-ray diffraction structures, which can be downloaded from the Protein DataBank. Each bead has the total charge of the residue it stands for. The beads are connected by springs and the number and strength of connections are adjusted to reproduce the vibrational normal modes of the protein.

I think that the combination of these two models could be used to improve greatly the results presented in the second part of this thesis, in particular concerning the 1D sliding of the protein. It could permit simulating the track of the protein on the double helix, and therefore should provide a more accurate 1D diffusion coefficient and sliding length. Most importantly, such a composite model would also be sufficiently fine-grained to model specific interactions and simulate how DNA-binding proteins stop on their targets. The key point is obviously to be able to make the link between the two models, that is, to define meaningful interactions between the

beads composing the DNA sequence and those standing for amino acids.

I have started implementing these models, choosing to begin with the TATA-binding protein. The TATA-binding protein is a transcription factor, which binds to a small DNA sequence that is rich in thymine and adenine. It then opens the double strand by bending it to an angle of 80°. The motivation of this choice is twofold. Firstly, the TATA-binding protein is a small protein (the C-terminal domain is composed of 180 amino acids), which structure has been determined at very high resolution (see, for example, references [163] and [164]) and is well conserved between different species. Secondly, this is the first protein, which, in eukaryotes, connects to DNA during the initiation process for transcription by the RNA polymerase. It is therefore not influenced by the presence of other proteins, which makes it a simple system to study.

I am confident that such mesoscopic dynamical models of DNA-protein interactions can help a lot in clearing these complex domains.

References:

1. F. Crick and J. D. Watson, *Nature* **171**, 737 (1953).
2. M. Levitt and A.Warshel, *Nature* **253**, 694 (1975).
3. J. F. Mercier, C. Kingsburry, G. W. Slater and B. Lafay, *Electrophoresis* **29**, 1264 (2008).
4. C. T. Zhang, *Phys. Rev. A.* **35**, 886 (1987).
5. M. Techera, L. L. Daemen and E. W. Prohofsky, *Phys. Rev. A* **40,** 6636 (1989).
6. T. Dauxois, M. Peyrard and A. R. Bishop, *Phys. Rev. E* **47,** R44 (1993).
7. M. Joyeux and S. Buyukdagli, *Phys. Rev. E* **72,** 051902 (2005).
8. M. Ptashme, *Nature* **214**, 232 (1967).
9. W. Gilbert and B. Muller Hill, *Proc. Nat. Acad. Sci. USA* **58**, 2415 (1967).
10. A. D. Riggs, S. Bourgeois and M. Cohn, *J. Mol. Biol.* **53,** 401 (1970).
11. O. G. Berg and P. H. von Hippel, *Annu. Rev. Biophys. Biophys. Chem.* **14**, 131 (1985).
12. H. Gutfreund, Kinetics for the Life Sciences, Cambridge University Press, Cambridge, (1995).
13. R. F. Bruinsma, *Physica A* **313**, 211 (2002).
14. R. J. Roberts, T. Vincze, J. Posfai and D. Macelis, *Nucleic Acids Res.* **35**, D269 (2007).
15. M. Coppey, O. Bénichou, R. Voiturez and M. Moreau, *Biophys. J.* **87**, 1640 (2004).
16. S. E. Halford and J. F. Marko, *Nucleic Acids Res.* **32**, 3040 (2004).
17. T. Hu, A. Y. Grosberg and B. I. Shklovskii, *Biophys. J.* **90**, 2731 (2006).
18. H. Merlitz, K. V. Klenin, C. X. Wu and J. Langowski, *J. Chem. Phys.* **125**, 014906 (2006).
19. M. Slutsky and L. A. Mirny, *Biophys. J.* **87**, 4021 (2004).
20. H. Jian, A. Vologodskii and T. Schlick, *Journal of Computational Physics* **136**, 168 (1997).

21. S. E. Halford, *Biochemical Society Transactions* **37**, 343 (2009).
22. M. D. Barkley, *Biochemistry* **20**, 3833 (1981).
23. R. B. Winter, O. G. Berg and P. H. von Hippel, *Biochemistry* **20**, 6961 (1981).
24. M. Hsieh and M. Brenowitz, *J. Biol. Chem*. **272**, 22092 (1997).
25. B. H. Zimm and J.R Bragg, *J. Chem. Phys*. **31**, 526 (1959).
26. B. H. Zimm, *J. Chem. Phys.* **33**, 1349 (1960).
27. D. Poland and H.A. Scheraga, *J. Chem. Phys.* **45**, 1456 (1966).
28. D. Poland and H. A. Scheraga, *J. Chem. Phys.* **45**, 1464 (1966).
29. E. Carlon, E. Orlandini and A. L. Stella, *Phys. Rev. Lett* **88**, 198101 (2002).
30. T. Garel and C. Monthus, *Journal of Statistical Mechanics – Theory and experiment*, Art. No. P06004 (2005).
31. D. Poland, *Biopolymers* **13**, 1859 (1974).
32. M. Fixman and J. J. Freire, *Biopolymers* **16**, 2603 (1977).
33. J. SantaLucia, *Proc. Natl. Acad. Sci USA* **95**, 1460 (1998).
34. J. SantaLucia and D. Hicks, *Annu. Rev. Biophys. Biomol. Struct*. **33**, 415 (2004).
35. R. Blossey and E. Carlon, *Phys. Rev. E*. **68**, 061911 (2003).
36. Y. Kafri, D. Mukamel and L. Peliti, *Phys. Rev. Lett*. **85**, 4988 (2000).
37. M. E. Fisher, *J. Chem. Phys*. **44**, 616 (1966).
38. R.D. Blake, J. W. Bizzaro, J.D. Blake, G. R. Day, S. G. Delcourt, J. Knowles, K. A. Marx and J. SantaLucia, *Bioinformatics* **15**, 370 (1999).
39. R. D. Blake and S.G. Delcourt, *Nucleic Acids Res.* **26**, 3323 (1998).
40. M. Joyeux, S. Buyukdagli and M. Sanrey, *Phys. Rev. E* **75**, 061914 (2007).
41. S. W. Englander, N. R. Kallenbach, A. J. Heeger, J. A. Krumhansl and S. Litwin, *Proc. Natl. Acad. Sci. USA* **77**, 7222 (1980).
42. S. Yomosa, *Phys. Rev. A* **30**, 474 (1984).
43. C. T. Zhang and K. C. Chou, *Chem. Phys.* **191**, 17 (1995).
44. G. Gaeta, *Phys. Lett. A* **168**, 383 (1992).
45. G. Gaeta, *Phys. Lett. A* **190**, 301 (1994).
46. G. Gaeta, *J. Biol. Phys*. **24**, 81 (1999).
47. G. Gaeta, *Phys. Rev. E.* **74**, 021921 (2006).
48. Y. Gao and E. W. Prohofsky, *J. Chem. Phys.* **80,** 2242 (1984).

49. S. Cocco, M. Barbi and M. Peyrard, *Phys. Lett. A* **253**, 161 (1999).
50. T. Dauxois and M. Peyrard, *The physics of solitons*, Nonlinear Science and Fluid Dynamics, Cambridge University Press, Cambridge (2006).
51. A. Campa and A. Giansanti, *Phys. Rev. E* **58,** 3585 (1998).
52. A. Campa and A. Giansanti, *J. Biol. Phys.* **24,** 141(1999).
53. S. Buyukdagli and M. Joyeux, *Phys. Rev. E* **77,** 031903 (2008).
54. Y.-L. Zhang, W-M Zheng, J-X Liu and Y.Z. Chen, *Phys. Rev. E* **56,** 7100 (1997).
55. M. Peyrard and A. R. Bishop, *Phys. Rev. Lett.* **62,** 2755 (1989).
56. S. Buyukdagli and M. Joyeux, *Phys. Rev. E* **73,** 051910 (2006).
57. S. Buyukdagli and M. Joyeux, *Phys. Rev. E* **76,** 021917 (2007).
58. S. Buyukdagli, M. Sanrey and M. Joyeux, *Chem. Phys. Lett.* **419,** 434 (2006).
59. D. De Luchi, C. Gouyette and J.A. Subirana, *Analytical Biochemistry* **322,** 279 (2003).
60. C. Danilowicz, Y. Kafri, R. S. Conroy, V. W. Coljee, J. Weeks and M. Prentiss, *Phys. Rev. Lett.* **93,** 078101 (2004).
61. C. H. Lee, C. Danilowicz, V. W. Coljee and M. Prentiss, *Eur. Phys. J. E* **19,** 339 (2006).
62. A. Brünger, C. B. Brooks and M. Karplus, *Chem. Phys. Lett.* **105,** 495 (1984).
63. T. Schneider and E. Stoll, *Phys. Rev. B* **22,** 5317 (1980).
64. S.G. Fisher and L.S. Lerman, *Cell* **16**, 191 (1979).
65. S. G. Fisher and L.S Lerman, *Proc. Natl. Acad. Sc. USA* **77**, 4420 (1980).
66. S. G. Fisher and L.S Lerman, *Proc. Natl. Acad. Sc. USA* **80,** 1579 (1983).
67. J. L. Viovy, *Rev. Modern Phys.* **72**, 813 (2000).
68. L.S Lerman, S.G. Fisher, I. Hurley, K. Silverstein and N. Lumelsky, *Ann. Rev. Biophys. Bioeng.* **13**, 399 (1984).
69. L.S. Lerman and K. Silverstein, *Methods Enzymol.* **155,** 482 (1987).
70. C. A. Malloff, R. C. Fernandez and W. L. Lam, *J. Mol. Biol.* **312**, 1 (2001).
71. C. A. Malloff, R. C. Fernandez, E. M. Dullaghan, R.W. Stokes and W.L. Lam, *Gene* **293**, 205 (2002).
72. E. M. Dullaghan, C. A. Malloff, A.H. Li, W.L. Lam and R. W. Stokes, *Microbiology* **148**, 3111 (2002).

73. G. Steger, *Nucleic Acids Res.* **22**, 2760 (1994).
74. S. Brossette and R.M. Wartell, *Nucleic Acids Res.* **22**, 4321 (1994).
75. R. L. Rill, A. Beheshti and D. H. Van Winkle, *Electrophoresis* **23**, 2710 (2002).
76. D. H. Van Winkle, A. Beheshti and R.L. Rill, *Electrophoresis* **23**, 15 (2002).
77. J. Zhu and R. M. Wartell, *Biochemistry* **36**, 15326 (1997).
78. R.M Wartell, S. Hosseini, S. Powell and J. Zhu, *J. Chromatogr. A* **806**, 169 (1998).
79. R. M. Meyers, T. Maniatis and L.S. Lerman, *Methods Enzymol.* **155**, 501 (1987).
80. N. Singh and Y. Singh, *Eur. Phys. J. E* **17,** 7 (2005).
81. S. Cocco, R. Monasson and J. F. Marko, *Proc. Natl. Acad. Sci. USA* **98,** 8608 (2001).
82. D. K. Lubensky and D. R. Nelson, *Phys. Rev. E* **65,** 031917 (2002).
83. E. Carlon, M. L. Malki and R. Blossey, *Phys. Rev. Lett.* **94,** 178101 (2005).
84. H. Urabe and Y. Tominaga, *J. Phys. Soc. Jpn* **50,** 3543 (1981).
85. H. Urabe, Y. Tominaga and K. Kubota, *J. Chem. Phys.* **78,** 5937 (1983).
86. E. W. Prohofsky, *Statistical mechanics and stability of macromolecules*, Cambridge University Press, Cambridge (1995).
87. S. Cocco and R. Monasson, *J. Chem. Phys.* **112,** 10017 (2000).
88. E. Protozanova, P. Yakovchuk and M. D. Frank-Kamenetskii, *J. Mol. Biol.* **342,** 775 (2004).
89. A. Krueger, E. Protozanova and M. D. Frank-Kamenetskii, *Biophys. J.* **90,** 3091 (2006).
90. T. Ambjornsson, S. K. Banik, O. Krichevsky and R. Metzler, *Phys. Rev. Lett.* **97,** 128105 (2006).
91. F. De los Santos, O. Al Hammal and M. A. Munoz, *Phys. Rev. E* **77,** 032901 (2008).
92. D. Cule and T. Hwa, *Phys. Rev. Lett.* **79,** 2375 (1997).
93. S. Buyukdagli and M. Joyeux, *Chem. Phys. Lett.* **484**, 315 (2010).
94. A. D. Riggs, H. Suzuki and S. Bourgeois, *J. Mol. Biol.* **48,** 67 (1970).
95. O. G. Berg, R. B. Winter and P. H. von Hippel, *Biochemistry* **20**, 6929 (1981).

96. G. Adam and M. Delbruck, *Reduction of dimensionality in biological diffusion processes*, in *Structural Chemistry and Molecular Biology*, Freeman, San Francisco (1968).
97. P. H. von Hippel and O. G. Berg, *J. Biol. Chem*. **264**, 675 (1989).
98. J. Gorman and E. C. Greene, *Nature structural and molecular biology* **15**, 768 (2008).
99. J. Elf, G. W. Li and X. S. Xie, *Science* **316**, 1191 (2007).
100. I. Bonnet, A. Biebricher, P.-L. Porté, C. Loverdo, O. Bénichou, R. Voituriez, C. Escudé, W. Wende, A. Pingoud and P. Desbiolles, *Nucleic Acids Res*. **36**, 4118 (2008).
101. B. Bagchi, P. C. Blainey and X. S. Xie, *J. Phys Chem.* B **112**, 6282 (2008).
102. P. C. Blainey, A. M. van Oijen, A. Banerjee, G. L. Verdine and X. S. Xie, *Proc. Nat. Acad. Sci. USA.* **103**, 5752 (2006).
103. A. Tafvizi, F. Huang, J. S. Leith, A. R. Fersht, L. A. Mirny and A. M. van Oijen, *Biophys. J.* **95**, L1 (2008).
104. Y. M. Wang, R. H. Austin and E. C. Cox, *Phys. Rev. Lett*. **97**, 048302 (2006).
105. J. H. Kim and R.G. Larson, *Nucleic Acids Research* **35**, 3848 (2007).
106. A. Granéli, C.C. Yeykal, R.B. Robertson and E.C. Greene, *Proc. Nat. Acad. Sci. USA* **103**, 1221 (2006).
107. Y. Wang, L. Guo, I. Golding, E. C. Cox and N. P. Ong, *Biophys. J.* **96**, 609 (2009).
108. D. M. Gowers and S. E. Halford, *EMBO Journal* **22**, 1410 (2003).
109. D. M. Gowers, G. G. Wilson and S. E. Halford, *Proc. Natl. Acad. Sc. USA* **102**, 15883 (2005).
110. M. V. Smoluchowski, *Z. Phys. Chem*. **92**, 129 (1917).
111. H. C. Berg, *Random walks in Biology*, Princeton University Press, Princeton (1993).
112. M. T. Record, W.T. Zhang and C.F. Anderson, *Adv. Prot. Chem*. **51**, 281 (1998).
113. P.J.W. Debye, *Trans. Electrochem. Soc*. **82**, 265 (1942).
114. U. Böhme and U. Scheler, *Chem. Phys. Lett*. **435**, 342 (2007).
115. B. Honig and A. Nicholls, *Science* **268**, 1144 (1995).

116. N. J. B. Green, and S.M. Pimblott, *J. Phys. Chem*. **93**, 5462 (1989).
117. A. Szabo, K. Schulten and Z. Schulten, *J. Chem. Phys*. **72**, 4350 (1980).
118. K. V. Klenin, H. Merlitz, J. Langowski and C. X. Wu, *Phys Rev Lett* **96**, 018104 (2006).
119. M. D. Donsker and S. R. S. Varadhan, *Commun. Pure Appl. Math.* **28**, 525 (1975).
120. F. Spitzer, *Probab. Theory Relat. Fields* **3**, 110 (1964).
121. A. M. Berezhkvoskii, Yu. A. Makhnovskii and R. A. Suris, *J. Stat. Phys*. **57**, 333 (1989).
122. A. J. Varshavsky, S. A. Nedopasov, V. V. Bakayev, T. G. Bakayeva and G. P. Georgiev, *Nucleic Acids Research* **4**, 2725 (1977).
123. M.D. Frank-Kamenetskii, A.V. Lukashin and V.V. Anshelevich, *J. Biomol. Struct. Dynam.* **2**, 1005 (1985).
124. D. Stigter, *Biopolymers* **16**, 1435 (1977).
125. A.V. Vologodskii and N.R. Cozzarelli, *Biopolymers* **35**, 289 (1995).
126. G. Arya, Q. Zhang and T. Schlick, *Biophys. J.* **91**, 133 (2006).
127. D.L. Ermak and J.A. McCammon, *J. Chem. Phys*. **69**, 1352 (1978).
128. J. Rotne and S. Prager, *J. Chem. Phys*. **50**, 4831 (1969).
129. J. G. de la Torre and V.A. Bloomfield, *Biopolymers* **16**, 1747 (1977).
130. B. Carrasco, J. G. de la Torre and P. Zipper, *Eur. Biophys. J.* **28**, 510 (1999).
131. M. Fixman, *Macromolecules* **19**, 1204 (1986).
132. R.M. Jendrejack, M.D. Graham and J.J. de Pablo, *J. Chem. Phys.* **113**, 2894 (2000).
133. C. Bustamante, M. Guthold, X. Zhu and G. Yang, *J. Biol. Chem*. **274**, 16665 (1999).
134. P. Etchegoin and M. Nöllmann, *J. Theor. Biol*. **220**, 233 (2003).
135. M. Barbi, C. Place, V. Popkov and M. Salerno, *Phys. Rev. E* **70**, 041901 (2004).
136. M. Barbi, C. Place, V. Popkov and M. Salerno, *J. Biol. Phys*. **30**, 203 (2004).
137. T. Hu and B.I. Shklovskii, *Phys. Rev. E* **74**, 021903 (2006).
138. R. Murugan, *Phys. Rev. E* **76**, 011901 (2007).
139. V. Dahirel, F. Paillusson, M. Jardat, M. Barbi and J. M. Victor, *Phys. Rev. Lett.* **102,** 228101 (2009).

140. S. Jones, P. van Heyningen, H. M. Berman and J. M. Thornton, *J. Mol. Biol*. **287**, 877 (1999).
141. O. Givaty and Y. Levy, *J. Mol. Biol*. **385**, 1087 (2009).
142. I. Golding and E. C. Cox, *Phys. Rev. Lett.* **96**, 098102 (2006).
143. G. Guigas and M. Weiss, *Biophys. J*. **94**, 90 (2008).
144. M. Wachsmuth, W. Waldeck and J. Langowski, *J. Mol. Biol*. **298**, 677 (2000).
145. R. Metzler and J. Klafter, *Phys. Rep.* **339**, 1 (2000).
146. M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan and P. Lu, *Science* **271**, 1247 (1996).
147. T. Schlick, *Molecular modeling and simulation – An interdisciplinary guide*, Springer, New York (2006).
148. J. G. Kirkwood and J. Riseman, *J. Chem. Phys*. **16**, 565 (1948).
149. J. M. Deutch and B. U. Felderhof, *J. Chem. Phys*. **59**, 1669 (1973)
150. H. L. Friedman, *J. Phys. Chem*. **70**, 3931 (1966).
151. Y. von Hansen, R.R. Netz and M. Hinczewski, *J. Chem. Phys*. **132,** 135103 (2010).
152. O. Seksek, J. Biwersi and A.S Verkman, *J. Cell. Biol.* **138**, 131 (1997).
153. J. Braga, J.M.P. Desterro and M. Carmo-Fonesca, *Mol. Biol. Cell*. **15**, 4749 (2004).
154. S. M. Gorsich, M. Wachsmuth, K. F. Toth, P. Lichter and K. Rippe, *J. Cell Sci*. **118**, 5825 (2005).
155. J. M. Schurr, *Biophys. Chem*. **9**, 413 (1975).
156. R. D. Phair and T. Misteli, *Nature* **404**, 604 (2000).
157. A. B. Houtsmuller and W. Vermeulen, *Histochem. Cell. Biol*. **115**, 13 (2001).
158. A. S. Verkman, *Trends Biochem. Sci*. **27**, 27-33 (2002).
159. R. D. Phair, P. Scaffidi, C. Elbi, J. Vacerova, A. Dey, K. Ozato, D. T. Brown, G. Hager, M. Bustin and T. Misteli, *Mol. Cell Biol*. **24**, 6393 (2004).
160. T. A. Knotts IV, N. Rathore, D. C. Schwartz and J. J. de Pablo, *J. Chem. Phys*. **126**, 084901 (2007).
161. E. J. Sambriski, D. C. Schwartz and J. J. De Pablo, *Biophysical Journal* **96**, 1675 (2009).

162. J. N. Stember and W. Wriggers, *J. Chem. Phys.* **131**, 074112 (2009).
163. D. B. Nikolov and S.K. Burley, *Nature Structural Biology* **1**, 621 (1994).
164. Y. Kim, J. H. Geiger, S. Hahn and P. B. Sigler, *Nature* **365**, 512 (1993).

Abstract

The first part of my thesis deals with the modelling of DNA denaturation. I first used a statistical model (Poland-Scheraga) to show that one can predict the final positions of the fragments during 2D electrophoresis assays with a precision greater than experimental uncertainties. Then, I improved a dynamical model developed in our group by showing how its parameters can be varied to get predictions in better agreement with experimental results that were not addressed until now, like mechanical unzipping, the evolution of the critical temperature with sequence length, and temperature resolution. In the second part of my thesis I present a dynamical model for non-specific DNA-protein interactions. This model is based on a previously developed “bead-spring” model for DNA with elastic, bending and electrostatic interactions, while I chose to model protein-DNA interactions through electrostatic and excluded-volume forces. For the protein, I used two simple coarse-grained models: I first described the protein as a single bead and then improved this description by using a set of thirteen interconnected beads. I studied the properties of this model using a Brownian dynamics algorithm that takes hydrodynamic interactions into account, and obtained results that essentially agree with experiments. For example, I showed that the protein samples DNA by a combination of 3D diffusion in the buffer and 1D sliding along the DNA chain. I have also showed that this process, which is known as facilitated diffusion, cannot accelerate DNA sampling by proteins as much as it is sometimes believed to do.

Keywords: 2D electrophoresis display, DNA denaturation, facilitated diffusion.

Résumé

La première partie de ma thèse porte sur la modélisation de la dénaturation de l'ADN. J'ai tout d'abord utilisé le modèle statistique de Poland-Scheraga pour montrer que, lors de l'électrophorèse 2D, on peut prédire les positions finales des fragments avec une précision meilleure que l'incertitude expérimentale. J'ai ensuite amélioré un modèle dynamique développé dans l'équipe en variant ses paramètres pour obtenir un meilleur accord avec des résultats expérimentaux nouveaux, tels la dénaturation mécanique, l'évolution de la température critique avec la longueur de la séquence, et la résolution en température. Dans la seconde partie de ce travail, je propose un modèle qui décrit les interactions non-spécifiques entre l'ADN et les protéines. Ce modèle est basé sur une description "billes et ressorts" déjà existante de l'ADN, qui inclut des interactions d'élongation, de pliage et électrostatiques, alors que je décris les interactions entre l'ADN et la protéine par des énergies électrostatiques et de volume exclu. Pour la protéine, j'ai tout d'abord considéré une simple bille, puis un réseau de treize billes interconnectées. J'ai étudié la dynamique de ce modèle en utilisant un algorithme de dynamique brownienne qui tient compte des interactions hydrodynamiques et montré qu'il donne des résultats en bon accord avec les expériences. J'ai par exemple observé que la protéine visite bien les différents sites de l'ADN par une succession de diffusion 3D et de glissement 1D le long de l'ADN. J'ai également montré que ce processus, appelé facilitated diffusion, ne peut pas accélérer beaucoup la vitesse de recherche de la protéine, contrairement à ce qui est parfois soutenu.

Mots-clés : électrophorèse en deux dimensions, dénaturation de l’ADN, diffusion facilitée.