UNIVERSITÉ DE LILLE
Faculté de Sciences et Technologies
Ecole Doctorale Sciences Pour l'Ingénieur

THÈSE

Pour obtenir le grade de

Docteur en Sciences de l'Université de Lille
Discipline: Micro-nano systèmes et Capteurs

Présentée par

Fabio LANDUZZI

PHÉNOMÈNES MOLÉCULAIRES DANS L'ENDOMMAGEMENT DE L'ADN PAR RAYONNEMENTS IONISANTS

Directeur de thèse: **Prof. Fabrizio CLERI**

Thèse soutenue le 14 décembre 2018 devant le jury composé de

M.me Elise DUMONT	Ecole Normale Supérieure de Lyon	Rapporteur
M. Jean-François ALLEMAND	Ecole Normale Supérieure de Paris	Rapporteur
M. Enrico CARLON	Katholieke Universiteit Leuven	Examinateur
M. Felix RITORT	Universidad de Barcelona	Examinateur
M. Ralf BLOSSEY	Université de Lille	Invité
M. Dominique COLLARD	CNRS Lille	Invité
M. Fabrizio CLERI	Université de Lille	Directeur de thèse
M. Pier Luca PALLA	Université de Lille	Co-encadrant

UNIVERSITY OF LILLE
Faculty of Sciences and Technologies
Doctoral School of Engineering Sciences

THESIS

To obtain the degree of

Doctor in Sciences for the University of Lille
Discipline: Micro-nanosystems and Detectors

Submitted by

Fabio LANDUZZI

MOLECULAR PHENOMENA IN DNA DAMAGE BY IONIZING RADIATION

Thesis Advisor: **Prof. Fabrizio CLERI**

Thesis defended on December 14, 2018 in front of the Committee composed by:

M.me Elise DUMONT	Ecole Normale Supérieure de Lyon	Rapporteur
M. Jean-François ALLEMAND	Ecole Normale Supérieure de Paris	Rapporteur
M. Enrico CARLON	Katholieke Universiteit Leuven	Examinateur
M. Felix RITORT	Universidad de Barcelona	Examinateur
M. Ralf BLOSSEY	Université de Lille	Invité
M. Dominique COLLARD	CNRS Lille	Invité
M. Fabrizio CLERI	Université de Lille	Directeur de thèse
M. Pier Luca PALLA	Université de Lille	Co-encadrant

ABSTRACT

This thesis is devoted to a combined theoretical and experimental investigation of the structure and dynamics of two common types of defects occurring in the DNA molecule, after chemical or radiation damage: *base mismatches* and *strand breaks*.

Base mismatches are local deviations from the ideal Watson-Crick pairing rules. Strand breaks are lesions to the DNA backbone, defined by the cleavage of the phosphodiester bond. Both such defects could occur either naturally, from imperfections in the cell process, or environment- and artificially-induced, such as in cancer radiotherapy.

In the experimental part of the thesis, we used single-molecule force spectroscopy performed by optical tweezers, to characterize DNA mismatches. Single base alterations were introduced in two types of short DNA hairpins, for which we measured the excess free energies, and deduced the characteristic kinetic signatures of the defect from the force-displacement plots. We demonstrated that it is possible to experimentally detect a single base-pair mismatch, working at the lower sensitivity limits of the technique. Experiments were accompanied by Molecular Dynamics (MD) all-atom simulations of the same molecules. We could confirm some experimental assumptions, obtain a microscopic description of the unfolding pathways, and demonstrate different degrees of cooperativity between the base pairs.

In the second part of the thesis, we designed structural models for the DNA strand-break defects, in the two key constitutive elements of the chromatin: the *DNA linker* and the *nucleosome*. We constructed a model for the linker, a 31-bp dsDNA random sequence, in which we introduced different cuts in the backbone, to simulate the presence of already formed single- or double-strand breaks (SSB and DSB), whose evolution was studied by MD. The results revealed a complex dynamics of the defect regions, with collective bond rearrangement dominating with respect to simple H-bond breaking. Such findings allow to establish necessary conditions for the events eventually leading to the ultimate fragmentation of the DNA molecule.

The nucleosome is a portion of dsDNA wound around a core of eight histone proteins. Using MD simulations of nucleosomes with DSBs inserted at various sites, we characterized the early stages of the evolution of this DNA lesion. Using different data analysis techniques we observe that DSB on the DNA filament tend to remain compact, with only the terminal bases interacting with histones, exposing key features of the DNA-protein interactions. By calculating the covariant mechanical stress, we demonstrate that this contribution is important in the coupled bending and torsional energy landscape, thus helping in the complex process of damage recognition.

RESUME

Cette thèse est consacrée à une étude théorique et expérimentale combinée de la structure et de la dynamique de deux types communs de défauts dans la molécule d'ADN, suivant des dégâts de radiation ou d'espèces chimiques: *mismatches de bases* et *cassures de brins*.

Les mismatches de base sont des déviations locaux de l'appariement idéale à la Watson-Crick. Les cassures sont des lésions au squelette phospate de l'ADN, définie par le clivage de la liaison phosphodiestere. Ces types de défauts entre autres pourraient arriver naturellement, par des imperfections dans les processus cellulaires, ou être induits par interaction avec l'environnement, et artificiellement comme dans la radiothérapie du cancer.

Dans la partie expérimentale de la thèse, nous avons utilisé la spectroscopie de force sur molécule unique par le biais de pinces optiques, au but de caractériser des mismatches d'ADN. Des mutations d'une base ont été inseré par synthèse dans deux types de "hairpin" d'ADN courts, pour lesquels nous avons mesuré l'excès d'énergie libre, et nous avons déduit les signatures cinétiques caractéristiques du défaut, en étudiant les courbes force/déplacement. Nous avons démontré qu'il est possible de détecter expérimentalement la présence d'un défaut de mismatch isolé, travaillant aux limites inférieures de sensibilité de la technique. Les expériences ont été accompagnées par des simulations de Dynamique Moléculaire (MD) tout-atomes, des mêmes molécules. Nous avons pu ainsi confirmer quelques suppositions expérimentales, obtenir une description microscopique des trajectoires de dépliement de l'hairpin, et démontrer des degrés différents de cooperativité entre les paires de bases.

Dans la deuxième partie de la thèse, nous avons realisé des modèles structurels pour les défauts cassure simple (SSB) et double-brin (DSB) d'ADN, dans les deux éléments constitutifs de la chromatine : le *linker* et le *nucleosome*. Nous avons construit un modèle pour le linker, un brin d'ADN de 31 paires de bases avec séquence aléatoire, dans lequel nous avons introduit des cassures différentes dans le squelette phosphate, pour simuler la présence de SSB et DSB déjà formés, dont l'évolution a été étudiée par MD. Les résultats ont révélé une dynamique complexe des régions de défaut, avec les réarrangements collectifs dominant par rapport aux simples clivages de liaisons hydrogène. Ces résultats nous permettent d'établir des conditions nécessaires pour la succession d'évenements menant finalement à la fracture de la molécule d'ADN.

Le nucleosome est un long brin de ADN enroulé autour d'un coeur de huit protéines, les histones. Par le biais de simulations MD de nucleosomes avec des DSB insérés aux divers sites, nous avons caractérisé les stades précoces de l'évolution de cette lésion d'ADN. En utilisant des techniques d'analyse de donnés très poussées, nous observons que le DSB a tendance à rester compact sur le filament d'ADN, seulement ses bases terminales interagissant avec les histones. Cela permet d'exposer des caractéristiques clés des interactions entre histones et ADN. En calculant le stress mécanique en formulation covariante, nous démontrons que cette contribution est importante dans le couplage entre courbure et torsion de la double hélice, ainsi aidant dans le processus de reconnaissance de dégâts par les protéines de réparation.

ACKNOWLEDGEMENTS

It is not easy to remember all the people that, during these three intense years, have contributed inside and out of the laboratory to the project that has brought to this thesis. Firstly, I would like to express my sincere gratitude to my supervisor Prof. Fabrizio Cleri for the continuous support of my Ph.D study and related research and to my co-tutor Dr. Pier Luca Palla. They supported me with their suggestions, their patience and motivation. They guidance helped me in all the time of research and in the writing of this thesis. The important people help you in the moment of difficulty and there have been many days with problems or unsatisfactory results, they have always helped me with their knowledge and with their calm support.

Besides them, I would like to thank the rest of my thesis committee: Prof. Elise Dumont, Prof. Jean-François Allemand, Prof. Enrico Carlon, Prof. Felix Ritort, Dr. Ralf Blossey and Prof. Dominique Collard, for their insightful comments and encouragement.

A special thanks goes to Prof. Felix Ritort and his team, that welcomed me in the Small BioSystem Lab sharing his knowledge on optical tweezers with a still inexperienced PhD student, and helped me in the difficult task of acquire dexterity in a field where I was a novice.

My sincere thanks goes to the Région Hauts-de-France (then Nord-Pas de Calais) and to the Président de l'Université de Lille, for the generous 3-year grant that funded my 2015-18 thesis project, without whom I would not have had this important opportunity of participating to the growth of scientific knowledge and my personal growth. I also thank the additional funding for my stay in Barcelona by the Ecole Doctorale SPI-072 and the Project EQUIPEX "Excelsior".

Regarding LyX: The LyX port was initially done by *Nicholas Mariette* in March 2009 and continued by *Ivo Pletikosić* in 2011. Thank you very much for your work and the contributions to the original style.

CONTENTS

ACRONYMS

MD Molecular Dynamics

MC Monte Carlo

ED Essential dynamics

PAC Principal component analysis

NN nearest-neighbors

FEL free-energy landscape

SMD Steered molecular dynamics

PME particle mesh Ewald method

RMSD root-mean-squared displacement

RMSF root-mean-squared fluctuation

HMM Hidden Markov model

TST Transition-state theory

FJC Freely-jointed chain

WLC Worm-like chain

OT Optical tweezers

MT Magnetic tweezers

AFM Atomic-force microscope

DNA Deoxyribonucleic acid

RNA Ribonucleic acid

ssDNA single-stranded DNA

dsDNA double-stranded DNA

SSB single-strand break

DSB souble-strand break

LET Linear Energy Transfer

UV Ultraviolet

IR Infrared

DDR DNA damage repair

HR homologous recombination

NHEJ non-homologous end-joining

MM mismatch

MMR mismatch repair

SSA single-strand annealing

ROS reactive oxygen species

FOREWORD

My original background is in theoretical physics, with a major in condensed matter physics and a specialization in molecular dynamics computer simulations. I obtained the *Laurea Magistrale* from the University of Bologna (Italy), under the guidance of Prof. L. Pasquini, in 2013. My last-year internship project was carried out in the ENEA Casaccia laboratories in Rome, under the guidance of Dr. M. Celino, and in the IEMN Lille, under the guidance of Prof. F. Cleri, with whom I worked for a few months on solid-state transformations in magnesium oxide.

The grant for the present PhD thesis was obtained in the framework of a research program on the molecular bases of DNA damage, under development at IEMN in Prof. Cleri's group, with the joint partial funding from the Region Nord-Pas de Calais (now Hauts-de-France) and the University of Lille. We decided to develop a research on DNA molecules including different types of atomic-scale defects as typically produced by ionizing radiation. The project was motivated by important open questions about the effects of radiation in cancer radiotherapy, in particular the mechanisms of damage that can lead to cancer cell arrest and death. Our research group is involved in a multi-scale modelling approach to this subject, ranging from theoretical models of multi-stable polymer chains, to Molecular Dynamics simulations of DNA superstructures, to the effects of radiation in the evolution of virtual cell aggregates, with the common background of establishing the physical conditions that determine the efficacy of radiation treatment.

The exploration of the biological consequences of radiation damage in cells starting from the physical-chemical modifications of DNA is an open field of research, which poses additional problems to the complexity of the subject, due to the radically different methods and languages used in the two fields of biology and physics. Based on my initial experience, in this wide research project I was initially in charge of using molecular dynamics, "all-atoms" simulations of DNA molecules including simulated radiation damage. After a first work dedicated to the study of defects in an isolated fragment of double-strand DNA, we initially tried to orient the study to the detailed chemistry of defect formation by radical attack; to this purpose, I spent some time in the Institute of Physics and Chemistry of Materials (IPCMS) in Strasbourg, with Dr. Mauro Boero, working on coupled quantum/classical simulations. However, we soon realized that such a type of study would have driven the project on a rather different path, so we dropped this line of research albeit reluctantly. The second part of the study was instead focused on enlarging the length- and time-scale of the system under study, and we turned our attention to radiation defects on the DNA in the nucleosome. This made the object of a very extensive computational study, in which record-

long computer simulations of the nucleosome were carried out, using a special grant of 6 million CPU hours on two of the largest supercomputers in France.

After the first year and half of my work, I was granted the opportunity to complement our investigation with experiments on single molecules by the optical tweezers technique, offered by the ongoing collaboration with the Small BioSystems Laboratory in Barcelona, directed by the professor Felix Ritort, a respected authority in the field. This collaboration was established with the long term goal of acquiring the competence for a future installation of the optical tweezers technology at the IEMN laboratory, to autonomously develop our research. This turn forced me to the endeavour of gaining experience in experimental methods, moreover within the short time limits imposed by the PhD schedule, and while continuing to follow the simulation part of the study at the same time. Coming from a theoretical physics background, setting my foot in the laboratory was definitely not a simple task. For this reason, the Chapters of this document relative to the experimental work might appear excessively descriptive, at times, and lacking some possibly important technicalities. I am confident that the effort and dedication that I have put in this complex part of the work could compensate for my inevitable deficiencies. I want to thank the Catalan laboratory, prof. Ritort and his research group, for the collaboration and the kind support provided during the twelve months I spent in learning and applying the mini-tweezer technology.

As a first summary of the work developed in this thesis, the following papers on the simulation results have been published:

- F. Landuzzi, P. L. Palla, F. Cleri, *Stability of radiation-damaged DNA after multiple strand breaks*, Phys. Chem. Chem. Phys. (2017) 19, 14641-14651
- F. Cleri, F. Landuzzi , R. Blossey, *Mechanical evolution of DNA double-strand breaks in the nucleosome.*, PLoS Comp Biol (2018) 14, e1006224

Moreover, the following paper on the experimental part developed in the third year is being submitted:

- F. Landuzzi, F. Cleri, I. Pastor, F. Ritort, *Detection of DNA mismatch defects by force spectroscopy on short hairpins* (tbd)

Partial results of the various Chapters have been presented in the following conferences and workshops:

- *eMRS*, Lille 2016, France (participation)
- 15^{E}*Journées de la matière condensé*, Bordeaux 2016, France (poster)
- *Physics and Biological Systems 2016*, Ecole Polytechnique, Palaiseau, France (poster)
- *DNA Mechanics and Dynamics*, Leuven 2017 (presentation)
- *SmallBioSystem Lab: group meeting*, Barcelona 2017 (presentation)
- *DNA Damage and Repair: Computation Meets Experiments*, Leiden (NL), 2017 (poster)

1 INTRODUCTION

1.1 THE DNA MACROMOLECULE

When in the 1886, Gregor Johann Mendel published his master work *Versuche über Pflanzenhybriden* (*Experiments on Plant Hybridization*) he deduced the role in the transmission of genetic traits, of those invisible entities that he called *invisible factors*, or with a modern language *genes*. It was only in the first half of the XX century, with the works of Avery, MacLeod and McCarty [11] and Hershey and Chase [64], that the deoxyribonucleic acid (DNA) would be identified as the molecule responsible for the transfer of inheritance patterns in living organism. In the early 1950's, Chargaff [28] discovered that the amount of guanine in DNA is equal to cytosine and the amount of adenine is equal to thymine, thus establishing the two basic pairing rules for the nucleobases. Then, it was necessary to wait until 1953, for James Watson, Francis Crick and Rosalind Franklin to identify the molecular structure of DNA.

The DNA is a heterogeneous polymer composed by monomers called *nucleotides*. Each nucleotide is composed by three sub-units: 1) the phosphate group; 2) the deoxyribose, a five-carbon sugar; and 3) a nitrogenous base. Due to the possible variants on the nitrogenous bases, there are four different types of nucleotides in the DNA:

- the two *purines*, adenine (A) and guanine (G);
- the two *pyrimidines*, cytosine (C) and thymine (T).

The sugar-phosphate units form an oriented chain, or *backbone*, in which the orientation is defined by the two carbon atoms in the sugar

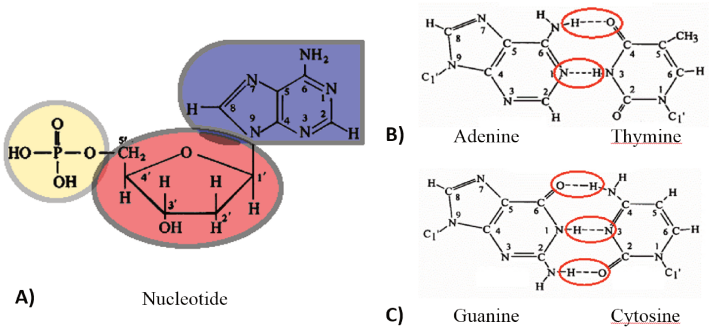


Figure 1: The nucleotide structure (a) in evidence the carbon position inside the sugar (in red) that determine the orientation of the backbone. At the C5′ it can be seen the phosphodiester bond with the phosphate group (yellow), and at the C1′ the bond with the adenine nitrogen base (in blue). On the right, the purine-pyrimidine pairing with the hydrogen-bond in evidence: (b) Adenine-Thymine with two H-bonds and (c) Guanine-Cytosine with three H-bonds.

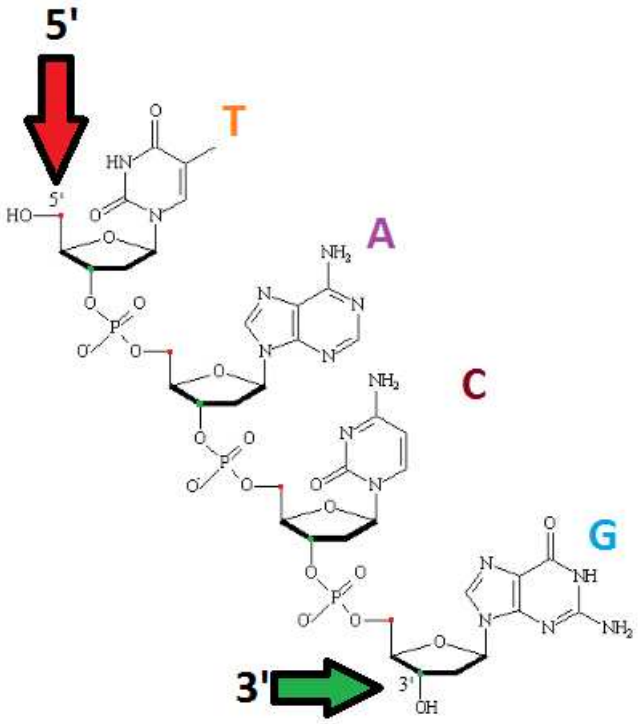


Figure 2: The termination 5′ and 3′ that are used to define the DNA chain orientation.

(5′ and 3′, respectively) where each phosphate group forms a covalent phosphodiester bond; each base, on the side opposite to the phosphate, forms a covalent bond with the first carbon of the corresponding sugar (Fig. 1 A). This why the orientation is the commonly described as 5′-termination to 3′-termination (Fig.2). Since the only difference along the DNA chain is in the nucleobases, it turns out that the whole genetic information is encoded in the sequence of the bases attached along the polymer strand.

The structure that Watson, Crick and Wilkins proposed for the DNA is a *double helix*, in which the nucleobases are coupled in complementary pairs attached on each side of the helix, and form hydrogen bonds across the helix axis. The two strands of complementary nucleotides have anti-parallel orientation, i.e., one runs from the 5′ to the 3′, and the other runs in the opposite direction. The standard, or Watson-Crick (W-C) pairing relations always involve one purine and one pyrimidine. In particular, adenine matches a thymine by forming two H-bonds, and guanine forms a stable pair with cytosine, forming three H-bonds (Fig. 1 B-C). Other pairing combinations, even if some of them are possible in principle, have a much smaller free energy of adhesion, and as a consequence they are less stable. Such weak pairings represent an error in the DNA code, and are called a *mismatch*.

When two complementary sequences are correctly matching, they form a ladder structure with the two backbones disposed laterally, and the bases paired pointing to the center (helical axis), like steps

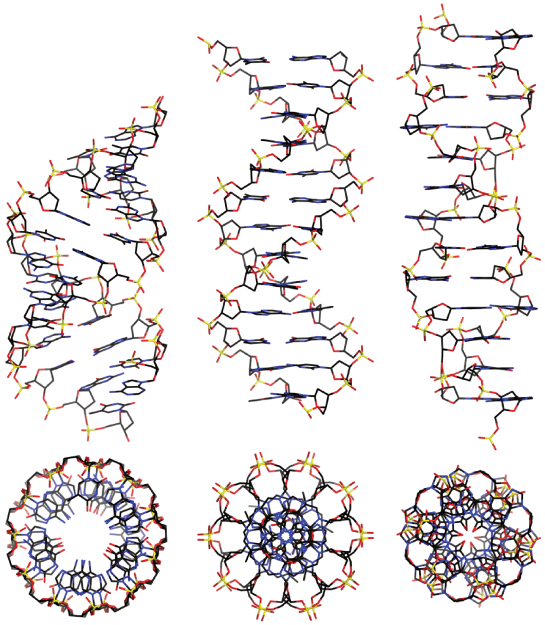


Figure 3: Three of the possible conformations of double stranded DNA: A-DNA (left), B-DNA (center) and Z-DNA (right). Image courtesy of Mauro Esguerroto reproduced under license CC-BY 4.0.

in a staircase. The double-helix folding is the result of the inner hydrophobic base-steps, which tend to reduce the spacing between them to avoid water contact; and of the rigid sugar-phosphate bond on the hydrophilic back-bone, which force the structure to writhe into a helix. Depending on the environmental conditions, the helix could assume different conformations. The predominant in the cell is the B-DNA (the one originally described by Watson and Crick), but also other forms such as A-DNA and Z-DNA have been experimentally observed. The B-DNA conformation is a right-handed helix with about ~10.4 base pairs per turn around the helical axis (that is, a base spacing of 3.4 Å along the backbone), a diameter of 2.0 nm, and an inclination of -1.2 Å between the DNA axis and the base pair plane; the helix pitch is not homogeneous, but two spacings are observed, namely a *major* and a *minor* groove, respectively of 22 Å and 12 Å width. The A-DNA conformation, usually observed under severe dehydration, is similar to the B but presents a higher number of base pairs per turn, ~11, while the Z-DNA is a left-handed helix, and it seems to appear only during the DNA transcription stage, likely as consequence of torsional stress (Fig 3). Other double strand conformations of the nucleotide polymer are possible, but there are no evidences of their existence in biological systems. Some evidence instead is found for triple-stranded and quadruplex conformations, in which 3 or 4 single strands of DNA combine together; these forms are peculiar to specific areas of the DNA, such as the telomeres.

The complementary base-pairing is fundamental for the inheritance of traits, or phenotypes, because the redundancy of information (each strand carries a copy of the same genetic information) allows the replication of information from one generation to the next, and the activation of cell repair mechanisms for the *damage* that naturally occurs

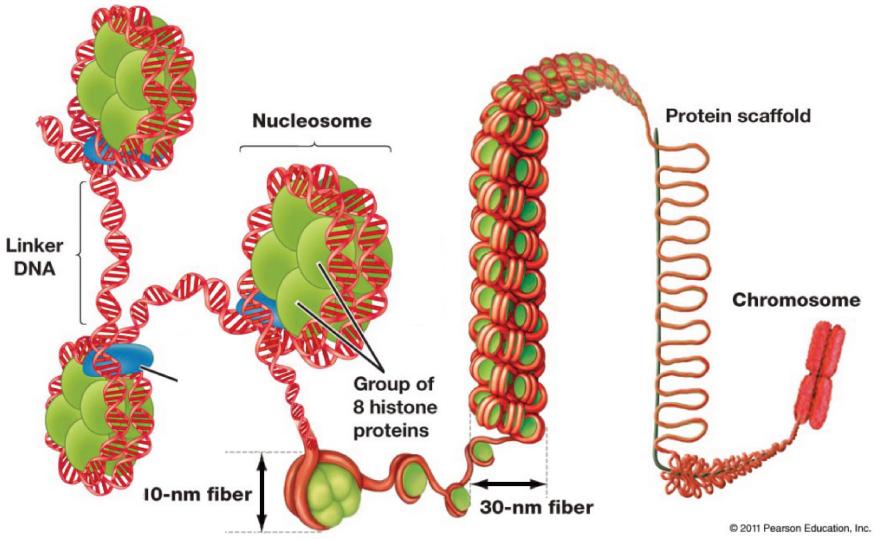


Figure 4: Hierarchical organization of the DNA inside the prokaryotic cells. The double-helix DNA is coil around the histone to form the nucleosome, then the *beads and string* sequence (*10-nm fiber*) is condensate in more compact structures that goes from the 30*nm chromatin fiber* to the *chromosome*. Image courtesy of Pearson Education Inc.

with high frequency during normal cell life. The double helix structure also poses a problem for the accessibility of the genetic information by transcription enzymes, that have to unwound the double-helical folded polymer in order to read the base sequence.

In human cells the filament of DNA, despite its nanometric diameter, has a macroscopic length of about 3 billion base pairs, corresponding to an extended contour length of about 1 m. Therefore, to be arranged within the cell nucleus of typically 5 µm size, it must be organized in a complex multi-scale structure (Figure 4), the *chromatin*. The fundamental unit of this macromolecule is the *nucleosome*, a spheroidal DNA-protein complex in which the DNA double-helix is wrapped around a units of 8 proteins, called *histones*. Nucleosomes are joined together in a beads-on-a-string sequence by segments of naked DNA, called *linker*. Because of the average diameter of the resulting filament, this lower level of chromatin organization is commonly referred to as the *"10-nm fiber"*.

Nuclear regions where the DNA is organized into higher-level chromatin structures are even more compact (the *"30-nm fiber"*), but also imply a reduced accessibility to the DNA information, so in *transcriptionally active* regions the chromatin is usually found in the less dense form, called *euchromatin*, as opposed to the denser, inactive regions of the cell nucleus, where it appears condensed as *heterochromatin*.

During the normal cell life, the genetic information carried by the DNA is continuously subject to mutation and damage due to various phenomena, such as: mistakes during the transcription, harmful chemicals introduced in the cells, or ionizing radiation. This change in information content could be the motor of evolution, by producing useful variations of some traits in the living organism, or as well be the cause of cell death and worse, of a fatal disease. Moreover, er-

rors in the repair process could generate in turn small insertion or deletion mutations, resulting in chromosomal rearrangements. The frequency of damage events is accounted between several thousand to a million per day in a human cell. Our particular interest will be now directed on the damage caused by ionizing radiation, mostly due to their widespread medical applications in cancer radiotherapy that originally motivated this thesis.

1.2 IRRADIATION: A NATURAL SOURCE OF DEFECTS, AND A MEDICAL TREATMENT

Nowadays, radiotherapy is one of the main treatments in use to kill or arrest cancer cells. It makes use of ionizing radiation to induce cell arrest, or apoptosis, a programmed mechanism of cell death, in cancer cells. *Ionizing radiation* is defined as any radiation (electromagnetic or particle emission) possessing enough energy to liberate electrons from atoms [26]. The lowest ionization energy is observed in Cs, with 3.89 eV. In the context of radiotherapy, such energy threshold commonly refers to the energy necessary to ionize the water molecule, that is 12.62 eV. Note that in medical practice, this threshold is more commonly taken to be 32 eV, the so called *W-value*, that is the mean energy necessary to form a ion pair plus some loss due to electronic excitations. The loss of an electron from atoms in a molecule can break covalent bonds, or produce a *radical*, a highly reactive chemical species capable of damaging nearby (biological) molecules. Both electromagnetic waves and high-energy charged particles can directly interact with the electronic cloud of atoms via electrostatic interaction, and cause the dislodging of electrons, whereas uncharged particles like neutrons can only indirectly interact, by causing atomic nucleus instability and a subsequent radioactive decay process.

High-energy radiation is naturally produced in the environment during the interaction of cosmic rays with the Earth's atmosphere, producing cascades of charged particles and energetic photons. A component of the environmental radiation exposure comes from the natural decay of rare elements in the Earth's crust, such as Radon-222 (which accounts for ~42% of the general population exposure to high-LET (Linear Energy Transfer) radiation. On the other hand, for the purposes of medical imaging or cancer radiotherapy, high-energy radiation can be artificially produced by particle beam accelerators, or artificial radioactive sources.

When ionizing radiations penetrate a biological tissue, their energy is released along the radiation path. Due to the different nature of radiations, the energy could be continuously released within some thickness (charged particles), or penetrate deep into the tissue and be delivered in discrete amounts (photons, neutrons). The quantity of energy released per unit distance, or LET, gives a qualitative measure of the danger related to the different kinds of radiation:

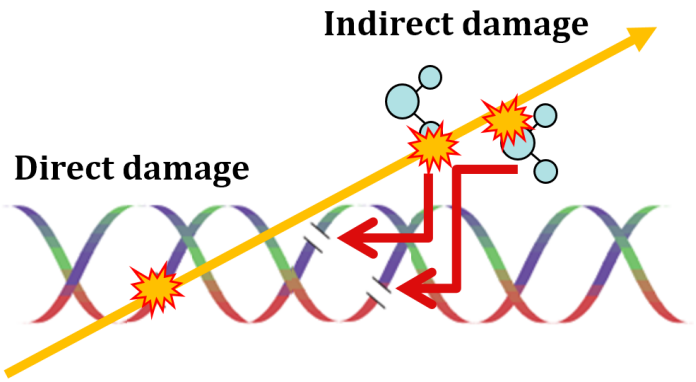


Figure 5: Schematic representation of the mechanism of direct and indirect damage in a DNA.

- low-LET radiations, such as photons or neutrons, can penetrate deeply into the tissue but usually release their energy in well isolated ionization and excitation events.
- high-LET radiations, such as protons or alpha particles, virtually ionize all the atoms along their path and as a consequence they have a small penetration depth, but can produce a dense shower of dangerous radicals.

The damaging mechanisms induced by ionizing radiation are customarily divided into two categories (Fig 5):

- *direct effects*, namely damage caused by the direct interaction of ionizing radiation with the atoms in the biological molecules. In particular, in the DNA such type of damage could cause structural changes.
- *indirect effects*, with a more complicated dynamics, implying firstly the ionization of molecules in the medium and the subsequent formation of free radicals, these latter diffusing and causing the chemical damage to other species.

When ionizing radiations pass through the cell they have a high probability to interact with the water molecules that are prevalent in the cell environment, and to induce *water radiolysis*. The radicals produced in this interaction can be absorbed by biomolecules, and modify their chemical structure, leading to various types of chemical damage. These indirect oxidative pathways are thought to be responsible for about 2/3 of the overall radiation damage to DNA (at least for low-LET, and notably UV-radiation), so they have been the subject of a considerable experimental and theoretical effort [23].

The direct effects usually produce isolated modifications. On the other hand, indirect effects produce often clustered defects due to the formation of a dense shower of radicals. Among all the biomolecules in the cell, the DNA is obviously the most sensible to (chemical and) radiation damage, because if not correctly repaired that would affect the irreplaceable blueprint of all cell activities.

Medical radiotherapy treatments exploit the vulnerability of cancer cells to ionizing radiation. The radiation beam is aimed at the tumoral region, previously identified and localized in the patient's tissues, with a time- and space-dependent dose profile, so as to maximize the damage to the cancer cells, and to minimize at the same time the (inevitable) damage to the nearby healthy cells. For future reference, the radiation dose is measured in units of Gray (Gy). Note however that the dose is not a measure of the biological damage imparted to the tissue. This latter depends on a complex network of interacting factors, and can be estimated by radiobiological models. The unit of biological damage is the Sievert (Sv), and depends on the nature of the radiation, for example 1 Sv is obtained from a dose of 1 Gy for high energy photons or electrons, or 0.5 Gy for a proton, or 0.05 Gy for an alpha particle.

1 Gy corresponding to the energy of 1 J delivered to 1 kg of matter

1.3 STRUCTURAL AND INFORMATION DEFECTS IN DNA

A consequence of the central role of DNA in cell life is the increasing interest on the genetic causes of many diseases. In fact, despite the huge size of the message encoded by the genome, even a single modification of the DNA sequence could induce dramatic consequences for the cell life. DNA does not directly act in cell life, but it is used as a template to produce RNA sequences (*transcription*) that are then *translated* into proteins, the building blocks of cell life. Proteins are long polymers made out of a set of twenty small constituents (*amino acids*), each amino acid being read from the genetic message as one group of three sequential nucleotides (a *codon*). Depending on their chemical composition (primary structure), and on their three-dimensional folding conformation (secondary and tertiary structure) proteins fulfil all the different functions in cell life.

Ionizing radiation can produce many different modifications in the DNA, such as cross-linking, base excision (AP-site), single and double strand breaks (SSB, DSB), oxidation, hydrolysis, methylation (Figure 6). A complex molecular machinery has been developed by the cell, to identify and repair these damages. It is worth noting the extremely conserved nature of these mechanisms, which are identical with little exceptions in all living organisms (note that the figure is taken from a plant biology journal), which also means that they were developed already very early in the evolution. The various lesions imparted by both natural, endogenous, or artificial actions, could result, possibly even after the cell repair mechanisms, in point mutations of the genetic message such as: *substitutions* (one base is incorrectly replaced with another); *insertions* (one nucleotide is added in one of the two strands during replication); *deletions* (one nucleotide is deleted from the sequence). At the cell scale, these point mutations could produce chromosomal rearrangements and genome instability.

A single addition/deletion of a base in the DNA sequence could results in a shift of the encoded message, and therefore a drastic change of all the following amino acids in the sequence that will affect the

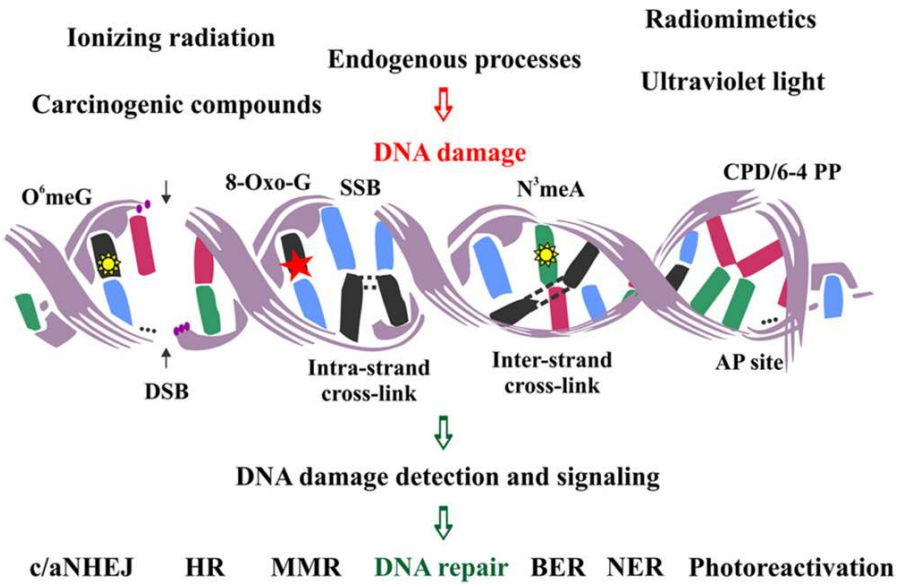


Figure 6: Schematic representation of the most important defects and repair mechanisms in DNA. (Reprinted under CC-BY-4.0 licence from ref.[105].)

whole protein composition. Even a single base substitution could generate a drastic change in cell functionality, as shown by the case of beta hemoglobin, for which even if one single glutamic acid (Glu) is mutated to valine (Val), in the sequence of 147 amino acids, the functionality of the entire blood cell is compromised (Sickle-Cell Anemia). Therefore, studying the DNA modifications and the repair processes has become a fundamental step, to understand the development and plan the cure of many diseases.

The formation of strand breaks is of special interest, since these defects open a physical cut in the DNA sequence, which requires a long succession of molecular events to be eventually repaired, and demands the right amount of information to properly execute the repair. Single-strand breaks (SSB) are rather easy to repair, since there is no lack of information (the other half of the DNA double helix is still intact), therefore the repair process is rapid and accurate, with typical repair times in the order of minutes. The double-strand break, or DSB, is especially lethal for the cell, since in this case the DNA is completely broken, and it can be repaired with great difficulty and with a high probability of errors. DSBs can persist in the cell nucleus ever several days after the damaging event. The presence of DSBs is at the origin of *chromosomal translocation*, that is the wrong rejoining of two parts of different chromosomes, whose ends had been cut open by different DSBs.

The repair processes for DSBs are classified in terms of:

(i) homologous recombination (HR);

(ii) non-homologous end-joining (NHEJ);

(iii) single-strand annealing (SSA);

(iv) microhomology-mediated end-joining (MMEJ);

The choice of a particular DSB-repair pathway depends on many criteria such as: type of damage, position of the damage in the nucleus, cell cycle phase. In mammalian cells, non-homologous end-joining (NHEJ) is the prevalent pathway for repairing DSB at any phase of the cell cycle, the broken DNA ends being simply pieced together in an efficient, but error-prone fashion [68, 90]. As already hinted, the damage is recognised by different "scout" proteins, such as the MRN complex, which also catalyse the recruitment of other proteins at the damage site, to start the rejoining mechanism. However, it is still not clear in which circumstances other repair pathways, such as HR or MMEJ, are activated. Notably, HR repairs DSBs in a generally error-free fashion, but since it requires an intact sister chromatid as a template, this mode of DSB repair only takes place in S/G2 phase. DDR is not immediate and likely both the position of the DSB in the cell nucleus and the complexity of the defect can lead to multi-stage repair kinetics. It is an open question as to which signals are actually being recognised since the earliest stages of the breakup process. Moreover, it is still open to debate whether radiation-induced DSBs develop immediately from ionization defects, or could rather evolve at later times, even minutes to hours after radiation exposure, as a result of extensive chemical processing of radiation-induced labile lesions.

In this thesis, we will focus our attention on two particular types of lesions in DNA: the **mismatch**, and the **strand break**. The first could result as a secondary effect of base pair alteration, the most common lesion due to reactive oxygen radicals, or from errors in the replication process. The second among the others is prominent, since physical breaks (single or double) in the backbone may lead to arrest in the cell cycle (upon halting the reproduction), and in the case of DSB ultimately to chromosome aberration, which ends up in cell death. As a rough estimate, often quoted in the medical literature, 1 Gy of low-LET ionizing radiation (photons, electrons) creates an average of 40-50 DSBs and about 1,000 SSB in each cell nucleus.

1.3.1 *Base-pair alteration*

Base-pair alteration is a modification in the molecular structure of the nitrogenous base, typically caused by a reactive radical.

A prominent example is the so-called 8-oxo-G lesion, a modified guanine which differs from ordinary guanine in that a H atom is replaced by a O at the C8 position, and the N7 nitrogen becomes protonated. Such defect can be formed by several reactive oxygen species (ROS) that are able to attack the guanine. In healthy human cells, the steady-state concentration of such defects is estimated to be around one 8-oxo-G per 10^6 guanines, obtained as a balance between rapid formation and robust removal by single-nucleotide replacement (~75% of occurrences) or by a slightly more laborious long-patch repair mechanisms (~25%, [42]). Before the repair, the mutated base maintains a similar structure to the original guanine, and continues

to form hydrogen bonds and correct stacking. However, if this defect is still present during the replication it can be wrongly paired with an adenine, resulting in a 8-oxo-G-A pair, thereby leading to a G-A mismatch; once the cell duplication is completed, this makes for a G-C to T-A mutation, since in the subsequent duplication the adenine will pair with a thymine, thus completely replacing the original base pair. Note that such a defect is completely undetectable by the repair enzymes, since it is a perfectly legitimate sequence; however, its consequences on protein synthesis can be disastrous.

1.3.2 *DNA mismatch*

A DNA mismatch is a structural defect occurring when two non-complementary bases are aligned in a sequence of duplex DNA [110]. Mismatches as are defined as *transduction* when formed by non complementary purine-pyrimidine bases, or *transversion* in the case of purine-purine or pyrimidine-pyrimidine pairs.

Compared to DNA strands with the canonical (Watson-Crick) pairing rules, mismatches are expected to produce alterations in the structure and stability of the DNA helix, especially in the proximity of the alteration site [119, 128, 156]. Mismatches (MM) can appear during replication of DNA,[57] heteroduplex formation [168], as well as by action of mutagenic chemicals, ionizing radiation, or spontaneous deamination [85]. MMs are efficiently corrected in DNA by mismatch repair (MMR) proteins, because failures in detecting or correcting the lesion could give rise to dangerous genetic mutations [85, 111]: in fact, MMs have been associated with 10-30% of spontaneous cancers in various tissues [81, 111]. In particular, G-A and G-T defects are of great interest to the cancer biology community, since such type of MM can be formed efficiently during oxidative stress, both by endogenous processes and following chemotherapy or radiotherapy.

Another common mismatch is the G-T, formed with high probability by polymerases such as β and Taq, during DNA replication, thus being about a thousandfold more frequent than other MMs; this is mainly due to its strong thermodynamic stability, which makes its identification by repair enzymes quite difficult [108].

Both the G-A and the G-T mismatch defects will be studied in the experimental part of this thesis.

1.3.3 *Single- and Double-Strand Breaks*

Strand breaks in DNA are defined by the cleavage of the phosphodiester bond that links two adjacent nucleotides. Such defects can be the result of endogenous cellular processes as replication or transcription of the genome, as well as of exposure to exogenous agents, such as radiation, oxidative and thermal stress, or certain chemicals [12, 129, 138]. It is assessed that isolated single-strand break (SSB) on the DNA can form by either a *direct* or *indirect* action of the radiation, or evolution of a damaged base into an apurinic/apyrimidinic (AP)

site, which is subsequently incised by an endonuclease enzyme. The natural formation rate of SSBs is quoted at about 55,000 per day per cell, or about one SSB per 100 kbp [155]; while high-energy ionising radiation, aimed at suppressing tumor cells in cancer radiotherapy, creates single-strand breaks at a rate of about 1,000 to 2,000 SSB/Gy [167] from both direct and indirect action.

During the evolution, highly efficient SSB repairing mechanism have been developed by the cells to compensate such a high rate of formation of these defects. Those mechanisms can conveniently use the undamaged strand as template, to restore the genetic information in a relatively short times, between 3 and 20 minutes [115]. On the other hand, the concurrent formation of strand breaks on both sides of the double helix compromises the possibility of using the complementary strand as a template, affecting the ability of the cell to restore the initial information, and can led to a fracture of the entire DNA in two separate fragments. This puts an accent on the importance of this type of defects in the development of the cancer, or lethal modifications for the cell life. Defects where the damage in the two strand happen at a distance that is less than one helical turn of the DNA (~10 base pairs) are classified as double-strand breaks (DSB).

In a simplistic vision, a DSB could be seen as resulting from two SSB formed independently. With such figures, the probability of formation of a DSB by two closely-spaced independent SSB would be rather small (as is indeed observed), and should grow with the square of the damage rate. While this is the special case for DNA exposed to chemical oxidants such as H_2O_2, the number of radiation-induced DSB is instead a *linear* function of the dose, with a rate of about 40 DSB/Gy,[37, 93] at least up to several hundred Gy, for low-LET radiation. This points to the fact that DSBs are unlikely to result from the statistical addition of two independently created SSBs (that would be, moreover, short-lived), but are produced at once by the radiation, which induces a dense swarm of ionisation products (mainly $OH^\bullet$ radicals and solvated e^-_{aq}) localized around the DNA fragment.

The severity of DSB lesions for cell life has forced the cells, during evolution, to find effective counter-measures. The *DNA damage response* (DDR) depend on the severity of the DSB (DSBs in clustered damage have smaller probability of being completely restored, than isolated DSB) but could be affected also by the particular stage of cell evolution during which the DSB has been detected. Choosing one of the different repair paths mentioned in the previous Section could result in different types of reparation. For example the *homologous recombination repair* is an error-free repair process, while the D-NHEJ repair path is faster but could cause translocation [68]. This will lead, in particular for complex damage, to the possibility that the repair process fails, or information is not completely restored and other modification occurs, such as a mismatch in the two strands.

In conclusion, a quite comprehensive understanding of the chemical mechanisms leading to phosphate bond cleavage is now available. However a few important steps still remains obscure, in the complex process of radiation damage to DNA. Among these, the de-

tailed mechanics of the transformation of the localized SSB and DSB damage, into a complete fracture of the molecule. In fact, even after the phosphate backbone is cut on both sides, a considerable binding energy from non-covalent interactions still remains to keep the fragments together: hydrogen bonds between the nucleotides, π-stacking interactions among the vertically piled aromatic cycles, electrostatic screening by the ions, are the main forces that are not immediately affected by the phosphate bond cleavage. For a quick comparison, the free energy of an isolated phosphodiester bond is estimated to be about 5.3 kcal/mol [43], while the residual free energy as deduced from DNA melting curves [130] is of the order of 1 kcal/mol per bp. Therefore, the two become comparable already for a DSB spaced by 3 or 4 base-pairs. So, in the second part of this thesis (based on molecular dynamics simulations) we decided to focus our attention on the single and double strand breaks in the DNA structure, because of their importance in cell life, either as product of endogenous cell mechanisms, or as a consequence of external cell damage.

1.4 STANDARD BIOLOGICAL METHODS FOR STUDYING DEFECTS ON DNA

When, due to the severity of the damage, the cell machinery fails in properly restoring the DNA information, there are three possible responses:

- *apoptotsis*, the cell is forced in a programmed death cycle when there is enough DNA damage to trigger the apoptotic signalling cascade;
- *senescence*, the cell ceases to divide, and enters in an irreversible dormant state;
- *neoplasia*, the cell begins uncontrolled division and, in the case of a malignant (invasive) evolution, can produce a cancer.

The analysis of DNA damage is essential to understand the transformation of a healthy cell to an apoptotic, senescent, or a cancer cell. Experimental methods to measure DNA damage in biology are based on the use of chemical manipulation techniques, and the quantification of damage can be done by optical microscopy or computer-automated counting.

The most commonly used methods for the SSB, DSB and mismatch detection in DNA are:

- **Single-cell gel electrophoresis assay**, (or "comet" assay), developed by Ostling, Johasson and Singh in the '80, it is a fluorescent microscopy-based method to detect various defects in the DNA structure, in particular SSBs and DSBs [31, 113]. Agarose-embedded cells are treated with solution to remove the cell environment, and form nucleoides containing supercoiled loops of DNA linked to the cell matrix. Exposed to an electric field, the

DNA tends to migrate inside the gel, but the supercoiled structure and the matrix links prevent the undamaged sequences to penetrate deep into the gel, while the damaged (broken into smaller pieces) sequences more easily unwind and migrate. As a consequence, the resulting migration path depends on the number of breaks in the DNA, and the proportion between the slowly migrating head of the path, and the diffusing tail, bears relation with the number of strand breaks in the sequence. The DNA is detected using fluorescent techniques, and it could be used to detect also the kind of defects present on the sample. The name *comet essay* originates by the particular shape of the head-to-tail migration path.

- **γ-H2AX and 53BP1 immunostaining**. The presence of DSBs in the chromatin of higher eukaryotic cells provoques a cascade reaction, which at some point leads to the phosphorylation of the histone H2AX variant into the γ-H2AX. This variant of the histone can be detected by immunofluorescence spectroscopy, as a marker in the studies of the DSBs. Automated counting can reveal the fraction of defects, and allows to follow the defect kinetics, by repeating the counting at regular intervals. A similar analysis can be done by tracking the protein 53BP1, which is early recruited at the damage site of SSBs.

- **Breaks Labeling In Situ and Sequencing** (BLISS). Cells or tissue sections are attached and fixed with formaldehyde onto a microscope slide or coverglass, which enables all the subsequent *in situ* reactions to be performed without centrifugations, thus minimizing the risk of introducing artificial DNA breaks and sample loss. DSB ends are blunted *in situ* and then ligated with a double-stranded DNA oligonucleotide adapter containing a unique molecular identifier. Following genomic DNA extraction, the portion of sequence immediately downstream to the tagged DSB is linearly amplified, and subsequently detected by PCR (protease chain reaction).

- **Bulky rhodium intercalators**, a method especially developed for targeting DNA mismatches. Some metal-intercalated molecular complexes have been found to selectively bind the thermodynamically destabilized, DNA-mismatch sites. Once bound to the mismatch, upon photoactivation, such metal complexes promote strand scission in the bases neighboring the mismatch. The fragments are then denatured and counted by standard polyamide gel electrophoresis (PAGE). Note that this is a very time-consuming technique, the synthesis of metal-intercalated complexes taking about a week, and the subsequent analysis a few days.

1.5 BIOPHYSICAL METHODS FOR THE EXPERIMENTAL STUDY OF DNA

The biophysical study of DNA started in the mid '50 with the outstanding X-rays analysis by Watson, Crick and Franklin, which led to the definitive assessment of the true structure of the DNA double-helix. Diffraction methods are still today largely used to determine structural features of biomolecules, together with methods based on nuclear spin resonance (NMR). In recent years, new sequencing methods based on the optical detection of fluorescently tagged DNA fragments have known a vast development (micro-arrays). Also DNA defects have been extensively studied and characterized by these methods. However all such methods are inherently *static*, i.e. they are all based on the determination of time-averaged properties of the molecules, such as averaged structures (X-rays, NMR) or their state of bonding (micro-arrays). If one is interested in the following the *dynamics* of the molecular system, notably its structural and chemical evolution in time following biochemical/biophysical interactions, it is necessary to resort to the most advanced developments of single-molecule dynamic force spectroscopy.

Single-molecule manipulation has known a huge development in the past years, and nowadays plays a central role in the understanding of many biological mechanism at the molecular scale. Such methods for example allow to investigate the mechanical properties of single polymers (DNA, RNA, or protein), individual chemical bonds, and interactions between biomolecules (such as DNA-protein mechanisms). In the study of DNA properties, given its peculiar filamentary structure, one extremity of the strand can be fixed to a support, and the other one attached to a force sensor; then, upon applying a controlled external force, the free-energy landscape of the molecule under various conditions can be explored. These methods go under the general name of dynamic force spectroscopy (DFS), and are used to probe single-molecule bond relationships between forces, lifetime, and chemistry. The understanding of these complex relationships will be treated with more detail in the Section 2.2 devoted to the theory of Bell and Evans. Recent studies demonstrated the possibility of detecting the presence of DNA mismatches with single-molecule force spectroscopy experiments, thus opening a new field in the understanding of DNA damage. Such techniques allow the real-time observation of the effects of a lesion on the molecule, and to compare the behaviors of the pristine sequence with the damaged one.

The most common methods to perform single-molecule biomolecular manipulation are: Atomic Force Microscopy (AFM), Optical Tweezers (OT) or Magnetic Tweezers (MT). A brief description of these techniques is given below.

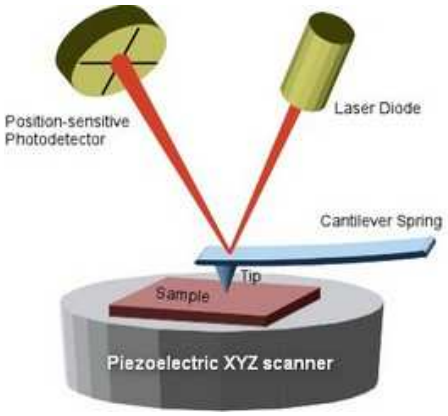


Figure 7: Schematic of the atomic-force microscope. The cantilever is deflected by the interaction with the sample and the position change is measured by the movement is detected by a photodetector.

1.5.1 *Atomic Force Microscopy*

The AFM is a development of the Scanning Tunneling Microscope (STM), based on the principle that a cantilever with an atomically-sharp tip can sense the roughness of the surface over which it is being dragged (Figure 7). The tip interacting with the surface deflects the cantilever, this deflection is measured and, knowing the equivalent stiffness of the instrument, it is possible to determine the strength of the interaction between the tip and the surface. Clearly, the same principle would apply if other media are interposed between the tip and the surface. Therefore, in single-molecule experiments the typical experimental setup is to coat a surface with the molecules under study, then use the cantilever tip to pick-up one of them, and record the force applied (displacement of the cantilever) during the removal of the tip at constant velocity from the surface. The adhesion to the tip could be aspecific, or specific if the extremities of the molecule are functionalized.This procedure allows the measurement of inter- and intra-molecular interaction forces at the piconewton-level. The range of force covered depends on the cantilever stiffness, which could reach the 1000 pN, while the spatial resolution is limited by the thermal fluctuations. For biological experiments at room temperature, by using the equipartition theorem with a cantilever stiffness in the range of ~100 pN/nm, one has $\Delta x = \sqrt{\delta x^2} = \sqrt{k_B T / k_{AFM}} \sim$ 0.1 nm, and for the force $\Delta f = \sqrt{\delta f^2} = \sqrt{k_{AFM} k_B T}$ ~10 pN. One main limitation of AFM in single-molecule experiments is the presence of uncontrolled interactions between the tip and the substrate, and its high stiffness compared to the OT an MT (See below), which therefore makes this method more suitable for strong molecular interactions.

1.5.2 *Magnetic Tweezers*

Magnetic tweezer experiments are based on the principle that a magnetic dipole, $\vec{\mu}$, immersed in magnetic field ($\vec{B}$) gradient experiences a

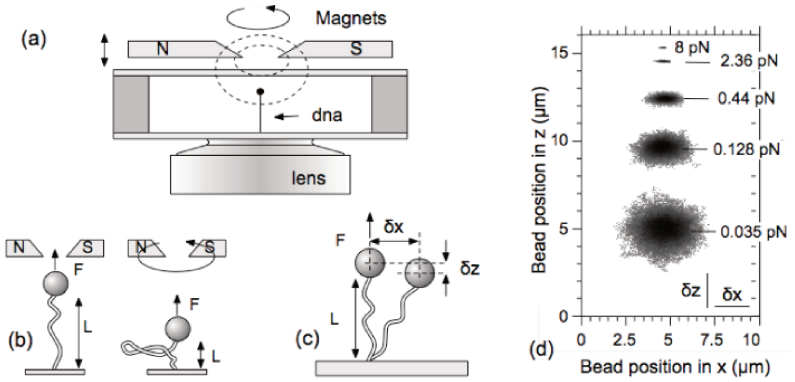


Figure 8: Magnetic tweezers scheme for polymer studies. a-b-c) The magnets generate a field that attracts the polarized bead, forcing the polymer to extend and rotate in response to the movement of the magnets. c-d) The bead attached to the polymer is forced by the magnetic field, and behaves like a damped pendulum, whose oscillation amplitude depends on the force exerted on the bead.

force $f = \vec{\mu} \cdot \nabla \vec{B}$. In a typical experiment setup (Figure 8), the molecule is attached to a magnetized bead and immersed in a strong magnetic field generated by two electromagnets, while the other extremity of the molecule is fixed to a surface. Because the magnetic interaction depends on the orientation of the dipole inside the field, moving the magnets this technique can be used not only to stretch the molecule but also to twist it. The measurement of the beads position is carried out by recording the beads position with a microscope objective and a CCD camera. The system is similar to a damped pendulum pulled by a force directed along the magnetic field gradient (say axis z). Then using the equipartition theorem is possible to relate the Brownian fluctuation of the bead to the force applied on it (for more details [149], [58]). The typical range of operating forces for this instrument is 10^{-2}-10 pN, since the magnetic trap stiffness is around 10^{-4} pN/nm, with a spatial resolution of 20 nm. The experiments are usually done fixing the position of the magnets and then recording the position. In this condition, even with such a large position fluctuation the magnetic field can be considered uniform, and the force exerted on the molecule is considered constant (at force clamp). Magnetic tweezers are one of the most important tools to perform single-molecule torque experiments on DNA and enzyme-DNA interaction. Despite such an advantage, the video detection prevents the direct measurement of very fast, or very small displacements; moreover, the large applied torque ($\sim 10^3$pNnm for a 1μm beads) limits the use of this feature, and requires special labeling of the molecule to directly measure the torsion.

1.5.3 *Optical Tweezers*

This method makes use of *optical traps*, based on the force generated by a focused laser beam acting on an object having an index of re-

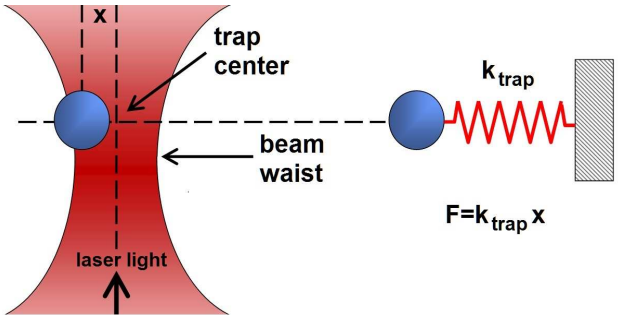


Figure 9: Optical tweezers schema for single molecule studies. The bead on the left, irradiated with a focused laser light tend to maintain its position centered in the focal point of the beam. The potential feels by the bead is analogous to the one feels by particle on the right, trapped in an harmonic potential with spring constant equivalent to the stiffness of the optical trap and the minimum localized its center. A molecule can be attached by one extreme ends to the beads, while the other is fixed, in and manipulated by moving the optical trap position. The applied forces and resulting displacements can be measured with pN- and nm-scale resolution.

fraction higher than that of the surrounding medium. If the system is correctly designed, a bead in the optical trap feels an attractive harmonic potential with minimum on the focal point of the beam (Figure 9). The momentum transferred from the light to the bead changes the direction of the beam, and by measuring the position and the intensity of the detected light it is possible to measure the force applied on the trapped object. In polymer studies, the typical trapped object is a micron-sized polystyrene or silica bead, which sets the range of forces explored at 0.1-100 pN. This method allows to measure at the same time force and elongation of the molecule, with a time resolution in the order of ~10 kHz. The most important limitations of the method are determined by the impossibility of measuring the molecule torsion, and the rather soft equivalent stiffness of the trap. More detail on this technique will be given in the Chapter 3, entirely dedicated to the Optical Tweezers experiments performed in the Small Systems Laboratory of the University of Barcelona, and in the Appendix A.

1.6 COMPUTER SIMULATION

All the above experimental techniques, even the most advanced single-molecule methods, encounter a limiting factor in the accessibility of the information on the molecule under study. If our objective is to study in molecular details the mechanics and kinetics of damaged DNA, one major limitation would be the timescale of the damaging process. Free radicals at the origin of the indirect damage have a typical lifetime of $\sim 10^{-9}$ s, and direct damage is even faster; subsequently, chemical defects evolve over typical time scales of microseconds. As said, optical traps can record data at a frequency of the order ~10 kHz, many orders of magnitude slower than the damaging process. Moreover, the direct information that can be obtained from the experiment

is relative to some global parameters, such as the force applied at the ends of the molecule, or the total extension of the polymer. Even the combination with fluorescent tagging techniques could add little information about the internal organization of the molecule during the damage evolution, or the relative position of interacting biomolecules.

To complement such limitations, computer-based molecular simulations can be introduced, in which the structure of the molecules of interest is described with a mathematical model, and its dynamic evolution under external perturbations can be studied in detail, even down to the single atomic position if desired. In this respect, there exist different modelling schemes that could be used, depending on the level of detail that one wants to attain in the molecular system description, each scheme having both advantages and intrinsic limitations:

- **Continuum Mechanics** describes the system as a continuous substance that completely fills the space that it occupies. A continuous body is one that can be continually sub-divided into infinitesimal elements, its properties being those of the bulk material, and not influenced by the molecular structure of the matter. The informations about the single atoms that compose the material are lost, and the microscopic information survives only in the form of local properties such as density, charge, polarizability, and so on. The equations describing the evolution of the system in this framework can be divided in two groups: *fundamental equations* (conservation of mass, conservation of charge, energy balance, linear and angular momentum balance), describing the physical laws which the body must obey independently from the material it is made of; and *constitutive equations*, that describe the relations between physical quantities in a specific material. The time and spatial evolution of the system is obtained from the analytical or, more often, numerical solution of partial differential equations for the fluid that describes the body.

- **Molecular Dynamics** describes the temporal evolution of a model system composed by point-like particles (the "atoms") obeying classical mechanics and interacting via effective potentials. The Newton's equations of motion for the ensemble of points are solved numerically using finite-difference methods. It is then possible to reconstruct the time-space trajectories with atomic precision, and the particle-particle interactions for each atom in the system; moreover, in systems that satisfy the statistical principle of *ergodicity* it is possible to obtain also a rich thermodynamic information (free energy, pressure isotherms, transport properties, thermal properties etc.). The main limitations of this method are imposed by the amount of computational resources necessary to solve the equations of motion, which limit the practical size of the system to a few million atoms at most, and to a few microseconds of simulation time; and by the reliability of the atomic interaction potentials ("classical" molecular dynam-

ics), which define the underlying dynamics of the system and its adherence to the experimental reality.

- ***Ab initio* quantum chemistry**, computational methods based on the solution of the quantum-mechanical Schrödinger equation for the electron cloud that determines the energy-minimizing position of the nuclei in the system, under various levels of approximation (Hartree-Fock, Born- Oppenheimer, Car-Parrinello, Bethe-Salpeter...). The atom-atom interaction forces are obtained as a result of the variation of the total energy landscape ("ab initio" molecular dynamics), and do not need to be a priori imposed in the form of empirical interatomic potentials. Such methods provide extremely detailed informations such as charge densities, bond formation and breaking, ground-state and excitation energies, and other properties of the system that do not depend on the fitting of empirical parameters. The main limitation of these techniques is imposed by the exceedingly large computational resources required to solve the Schrödinger equation even for a small ensemble of atoms, thus limiting the system sizes to a few hundreds of atoms and the time evolution to a few nanoseconds at most.

In the studies performed in this thesis, we will be chiefly interested in understanding the structural modifications and the consequent changes in mechanical properties of the DNA after the formation of such defects as mismatches and strand-breaks. Simple continuum-mechanics models of DNA described as a continuous polymer will be discussed in Section 2.1, as they will be crucial in the analysis and interpretation of the results of optical tweezer experiments.

On the other hand, the continuum-mechanics approach is not the best suited to describe the dynamical effects at the molecular scale, since the formation of a DNA defect represents a discontinuity in the structure, whose properties are not well defined in a continuum model. These phenomena will be rather treated by means of large-scale and extended-time classical Molecular Dynamics simulations, in which individual DNA defects can be explicitly followed and their microscopic evolution can be studied. However, considering the temporal limitations of a PhD thesis, it was decided to neglect the very early stages of radiation-DNA interaction and the subsequent radiochemical evolution. Such events occur on extremely rapid time-scales, compared with the typical time-scale accessible to classical molecular dynamics, and would require a detailed and separate study. Therefore we bypassed the dynamics of defect formation upon interaction with the ionizing radiation, thus ruling out *ab initio* quantum chemistry and electronic structure methods. Motivated by the possibility of a more direct theory-experiment comparison, we rather decided to focus on the mechanical consequences that follow the defect formation, with the possible evolution into base mismatches or even complete fracture of the DNA molecule. This choice implies that defects in the DNA structure will be always introduced by construction (both in the computer simulations and in the experiments), by choosing an initial

chemical form as the putative result of a preexisting interaction with radiation.

1.7 THE MOLECULAR SYSTEMS OF INTEREST

All the molecular systems object of study on this thesis are based on different chemical and structural arrangements of DNA fragments. For the molecular simulations, we always adopted all-atoms description of the system with the molecular interaction force field *charmm-27*. The three systems we have studied are:

- **DNA hairpins**, formed by a single strand of DNA composed by 82 bases. Due to its self-complementary pattern, the 24 bases in the central region are replied on itself and form a dsDNA of 10 base-pairs with an unpaired loop of 4 bases at one end. The two terminal segments of the strand at the opposite end are complemented with splint sequences, to form two double stranded DNA handles of 29 bps each. This same hairpin structure has been used as the basis also for the experimental studies by the optical tweezer technique. A more detailed description of the DNA hairpin system is given in Section 2.3, with the molecular dynamics model used in the simulations (Chapter 4), and a complete information about the defects introduced in the structure to perform the optical tweezer experiments in Sections 3.5 and 3.6.

- **DNA linker**, formed by a dsDNA sequence of variable length that connects two nucleosomes in the 10-nm chromatin structure (Figure 10). On this simple structure, elastically clamped at its ends to reproduce the chromatin background, we introduced different types of lesions, notability a single strand break and three different double-strand breaks. The models, the techniques used in molecular dynamics simulation and the results obtained are described in Chapter 6.

- **Nucleosome**, a complex supramolecular structure composed by a protein complex of 8 histones (pairs of H2A, H2B, H3, H4), and a long stretch of dsDNA winding around it (see again Fig. 10). Building on the results obtained in the study of linker-DNA, we introduced a DSB lesion in different positions of the nucleosomal DNA strand, and studied its characteristic structure and mechanical response in Chapter 7. Notably, the molecular simulations performed on massively parallel supercomputers (provided by French national Centres) for this part of the study are close to the world records of size-time extension for DNA simulations.

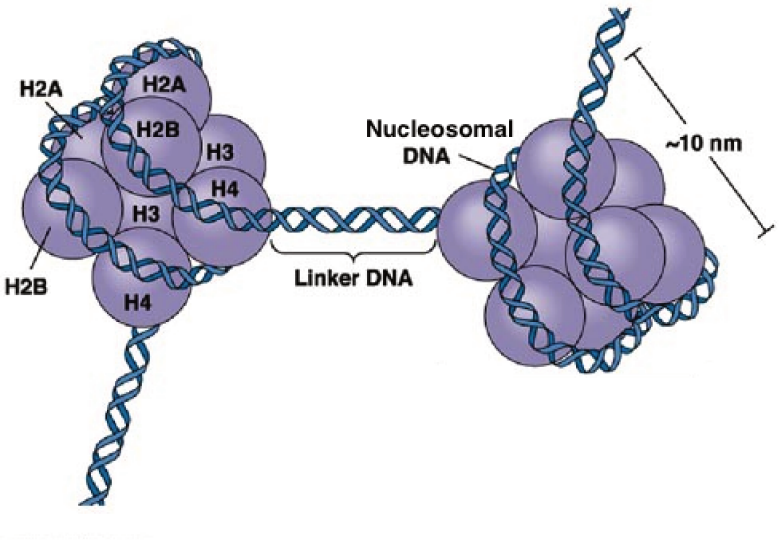


Figure 10: Schematic representation of a fragment of 10-nm chromatin fiber, showing the DNA linker connecting two nucleosomes.

1.8 LAYOUT OF THE THESIS

In the following Chapter 2 we will firstly describe the modeling of DNA by continuum mechanics, introducing the well-known models of the *freely-jointed chain* and *worm-like chain* from polymer physics. Together with other theoretical tools from statistical mechanics, such as the transition-state theory (also described in these chapters), such a theoretical apparatus will be used to interpret the hairpin folding/unfolding experiments described in Chapter 3, performed at the Small BioSystems Laboratory in Barcelona under the supervision of professor Felix Ritort.

However, it should be noted that this first part chronologically represents the third year of the thesis. It was decided to put this part before the Chapters describing the work performed in the first two years, because the hairpin is a structurally much simpler system, compared to both the full dsDNA and the nucleosome; moreover also the base-pair mismatch is a "simpler" defect, compared to the single and double-strand breaks described in Chapters 6 and 7. For the same reason of providing a better logical flow of the text, Chapter 5 describing the hairpin molecular simulations, which is the very last work developed at the end of the third year, for the sake of providing a much needed comparison with the experimental results, comes before the two Chapters 6 and 7, on dsDNA and nucleosome molecular simulations. It is hoped that such an arrangement, although not respecting the chronological development of the work, could indeed facilitate the reading of this document.

2

THE DNA AS A CONTINUOUS POLYMER

Mathematical models have played an important role in polymer physics since its early developments. Because of their highly repetitive structure, polymers have been often described in rather simple terms: their properties may be reduced to a small number of parameters describing the individual monomers, from which global continuum mechanics relations could be obtained, e.g. for the size dependence of volume, average conformation, elasticity, and so on. The DNA macromolecule, with its four bases repeated over long stretches, is a typical example of heterogeneous polymer, for which all the mathematical apparatus of polymer physics could be adapted. In the first part of this Chapter (Section 2.1) we will describe the models that are most relevant for the present study, namely the *freely-jointed chain* and the *worm-like chain*. Such a description of the DNA molecule as a continuous object, although ignoring the atomistic details, will be crucial in the interpretation of the experiments, which are carried out over timescales exceedingly long compared to the molecular ones, thus averaging out many molecular details.

This class of models does not support the explicit introduction of defects in the continuous structure of the polymer. Therefore, a different set of tools must be introduced in order to analyse the single-molecule experiments in which the DNA chemical configuration will be modified. The statistical mechanics description of the association/dissociation reaction is embodied in the *reaction rate theory*, which can be traced back to the early works of Van't Hoff and Arrhenius, and was more completely developed after the '30s, principally by Smoluchowski, Eyring and Kramers [62]. In the second part of this Chapter (Section 2.2), we will describe the latest developments of this formalism due to Bell, Evans and co-workers. Finally, in Sections 2.3 and 2.4 we will describe how these can be applied to the interpretation of single-molecule experiments.

2.1 CONTINUOUS POLYMER MODELS

Polymers are molecules composed by repetitive units called *monomers*. In biological molecules, like DNA, RNA or proteins (polypeptides), the monomers are covalently linked (*polymerization*) in long chains. All the classical developments of polymer models (e.g., Flory-Huggins, Rouse-Zimm, De Gennes reptation model) always considered the polymer chain to be a long, non-branched continuous filament. While some proteins can display a multi-branched structure, and branch-points can be created in nucleic acids during processes generating sequence rearrangements, such as homologous or site-specific DNA recombination, or by the secondary and tertiary folding of RNA, we

will restrict the analysis to the traditional models of non-branched chains. The conformation that those chains assume in solution is the result of the balance between local interactions, both short and long ranged, and entropy costs.

Different mathematical models have been introduced to describe the behaviour of such polymer chains in solution. The **Kratky-Porod** model describes the polymer as a homogeneous chain of rigid segments of length b, whose energy is determined only by the orientation of successive elements through a *bending modulus* parameter, B. Therefore, the partition function for the chain under an external force $\vec{f}$ is:

$$\begin{aligned} Z_{KP} &= \int d\Omega^N e^{\beta H_{KP}} = \\ &= \int d\Omega^N \exp\left[-\beta\left(\frac{B}{b}\sum_{i=1}^{N-1}\hat{t}_{i+1}\cdot\hat{t}_i + \sum_{i=1}^{N}\vec{t}_i\cdot\vec{f}\right)\right] \end{aligned} \quad (2.1)$$

where $\{\Omega\}_{i=1\ldots N}$ are the solid-angle coordinates that describe the orientation $\hat{t}_i$ of each element. Unfortunately, despite its apparent simplicity this model has no analytical solution for the relation between force and extension. However, the problem could be simplified for some limiting cases. In particular, we are interested in two such cases because they are important to describe the hairpin in optical-tweezer experiments: the simplest versions of the model are the so called "freely-jointed chain" (FJC) and the "worm-like chain" (WLC). These are rather ideal models of the polymer, for example none of them takes into account the intrinsic torsion of the DNA onto itself. This is not considered a strong limitation in optical-tweezer experiments, since in this case it is not possible to directly control the polymer torsion.

2.1.1 *Freely-Jointed Chain Model*

In the FJC model the elements that constitute the chain are considered as identical rigid segments, uncorrelated and free to move in every direction without paying an energy cost (Figure 11). Therefore, the only contributions to the free energy comes from the entropic term and the work done on the system by external forces. As said, this model is one limiting case of the more general Kratky-Porod model [149] for $B \to 0$, but unlike it, the FJC has an analytic solution. The total energy H_{FJC} of the system is given only by the work that an external force $\vec{f}$ has to do on the chain, to develop a certain extension x along the direction identified by the force vector.

The total extension is the sum of the contribution from each single monomer, then the energy coincides with the work done by the external force:

$$H_{FJC} = W_{ext} = -\vec{f}\cdot\vec{R} = -bf\sum_{n=1}^{N}\cos\theta_n \quad (2.2)$$

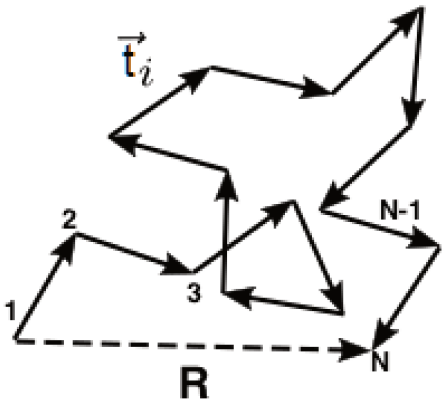


Figure 11: Freely-jointed chain. Scheme of a polymer described as a sequence of rigid monomers of fixed length, attached by their ends into a chain of elements with uncorrelated orientation. Bending two successive monomers by any angle θ has zero energy cost.

where b is the monomer length, N is the total number of monomers, and θ_n is the angle between the n-th monomer direction (defined by the vector distance between its two ends) and the direction defined by the force vector. With R the end-to-end distance, the above formula defines the energy as a function of the apparent size of the polymer, given by the distance between its opposite ends (the monomers 1 and N in Fig. 11). By exploiting the identity of the monomers, the partition function is obtained as:

$$\begin{aligned} Z_{FJC} &= \left[\int_{-\pi}^{\pi} d\theta \int_{0}^{2\pi} d\phi \sin\theta e^{\beta b f \cos\theta}\right]^N \\ &= \left[\frac{2\pi k_B T}{fb} \sinh\frac{fb}{k_B T}\right]^N \end{aligned} \tag{2.3}$$

where $\Omega = (\theta, \phi)$ is the solid angle, k_B the Boltzmann constant, T the system temperature, and $\beta = (k_B T)^{-1}$. We note that the monomer length b can be identified with the physical size of the chemical unit, or with the effective size of the irreducible rigid unit of an ideal chain equivalent to the physical polymer; in this case b is called the *Kuhn length*.

The force-extension relation can be obtained from the partition function, as a measure of the average value of the end-to-end distance:

$$\begin{aligned} x(f) &= \langle R \rangle = k_B T \frac{\partial \log Z_{FJC}}{\partial f} \\ &= Nb\left(\coth\frac{fb}{k_B T} - \frac{k_B T}{fb}\right) \end{aligned} \tag{2.4}$$

In the limiting case where the chain is constituted by only one monomer, $N = 1$, this solution describes the force needed to orient a rigid segment along a specific direction. Based on this latter deduction, we will later on use this model to describe the contribution to the variation of free-energy due to the orientation of the last hairpin bond along the pulling axis.

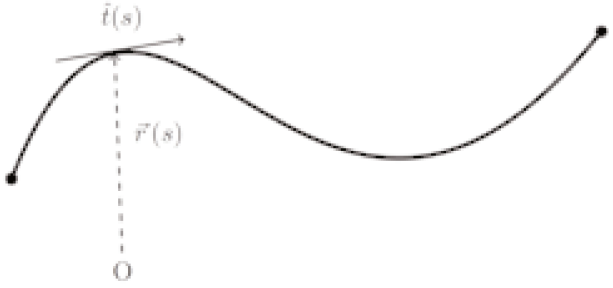


Figure 12: Worm-like chain. Scheme of a polymer described as a continuous sequence of point monomers of zero length, attached into a chain of elements with strongly correlated orientation. Bending two successive monomers by an angle θ costs an energy proportional to the bending modulus K of the filament.

The variation of free-energy due to the work of orientation of a rigid segment along the direction defined by the external force $\vec{f}$ is:

$$G_{orient}(f) = -k_B T \log\left[\frac{k_B T}{fb}\sinh\frac{fb}{k_B T}\right] \tag{2.5}$$

The FJC model can reproduce the observed linear force-extension relation of DNA at low stretching forces, and the observed saturation at high force. However, it fails in providing an accurate overall description of the entire DNA mechanical response.

2.1.2 *Worm-Like Chain Model*

The physical structure of DNA is far from being a sequence of uncorrelated, rigid monomers. Therefore it is not surprising that the simple FJC model does not correctly represent the whole behaviour of DNA. The best fit of the Kuhn length to force-extension data gives a value of $b \sim 100$ nm, orders of magnitude larger than the physical inter-nucleotide distance of 0.34 nm, a clear indication that more than something is wrong. It is therefore necessary to introduce a model that could take into account the energy needed to bend the polymer chain. In fact, the DNA has a rather large bending stiffness, which means that successive nucleotide pairs tend to point in more or less the same direction, the variation of the angle θ being relatively slow on the lengh scale of 0.34 nm.

The above cited Kratky-Porod model adds a term to the system energy in this direction:

$$H_{K-P} = -\frac{B}{b}\sum_{n=1}^{N-1}(\vec{t_{n+1}}\cdot\vec{t_n}) + H_{FJC} \tag{2.6}$$

This formula, in the limit where the segment length tends to zero (*continuous limit*) becomes the energy of the so-called worm-like chain (Figure 12):

$$H_{WLC} = \frac{\xi k_B T}{2} \int_0^{L_0} \left[\left(\frac{d\vec{t}}{ds} \right)^2 - \frac{|\vec{f}|}{k_B T} \cos\theta(s) \right] ds \tag{2.7}$$

The discrete variables $\vec{t}_n \to \vec{t}(s)$, $\theta_n \to \theta(s)$ are now functions of the curvilinear coordinate s along the chain *contour length* $L_0 = bN$. The generally temperature-dependent quantity $\xi = \frac{B}{k_B T}$ is defined as the *persistence length*, and is related to the decay of the angular correlation exponential:

$$\langle (\vec{t}(s_0 + s) \cdot \vec{t}(s_0)) \rangle = e^{-s/\xi} \tag{2.8}$$

There is no analytical expression for the force-extension curve of the WLC model. It is however possible to interpolate an approximate solution [106]:

$$f_{WLC}(x) = \langle f \rangle = \frac{k_B T}{4\xi} \left[\frac{1}{(1 - x/L_0)^2} - 1 + \left(\frac{4x}{L_0} \right) + 4 \sum_{n=2}^{7} a_n \left(\frac{x}{L_0} \right)^n \right] \tag{2.9}$$

The coefficients $\{a_n\}_{n=2\ldots7}$ are higher-order corrections to the formula; moreover, it is possible to introduce an enthalpic correction taking into account the extension of the contour length (internal stretching of the monomers) at high values of force. However, for the description of ssDNA in the hairpin open-state that will be used later on, both corrections can be generally ignored. The variation of free-energy needed to stretch the WLC along the direction defined by the external force vector $\vec{f}$ is:

$$\begin{aligned} G_{WLC}(f) &= \int_0^f x(f')df' = f\bar{x} - \int_{x_0}^{x_f} f(x')dx' \\ &= f\bar{x} - \frac{k_B T}{4} \frac{L_0}{\xi_T} \left(\frac{\bar{x}}{L_0} \right)^2 \left[\frac{1}{1 - \bar{x}/L_0} + 2 \right] \end{aligned} \tag{2.10}$$

where $\bar{x}$ is the value of x for which the Eq.(2.9) is equal to the actual value f of force, and the corrections of the coefficients $\{a_n\}_{n=2\ldots7}$ are ignored.

2.2 REACTION-RATE THEORY

In biology, non-covalent interactions between macromolecules govern structural cohesion, and determine the biomolecule folding process. Despite their paramount importance in cell life, such forces are relatively weak bonds with short lifetime, and even small stresses could prevent the bond formation or reverse the adhesion. The relation between force-lifetime-chemistry is the key to understand bond formation and disruption at molecular level in the biological context.

Single-molecule force spectroscopy has become a powerful tool to investigate the relation between force and lifetime, and the reaction-rate theory provides the theoretical framework to interpret the observations, and correlate them with the underlying chemistry.

Rate theory is the discipline that tries to explain chemical reactions with the physical formalism of the transition across metastable states. It was introduced by the works of Arrhenius and Van't Hoff in the late 1880s, while there were trying to describe the chemical transition between two states N and U

$$N \underset{k^-}{\overset{k^+}{\rightleftarrows}} U$$

Knowing the *kinetic rate* of the *forward* and *inverse* reaction k^+, k^- it is possible to evolve the population of the reactant n_N and product n_U, by solving the coupled differential equations:

$$\begin{cases} dn_N/dt &= -k^+ n_N + k^- n_U \\ dn_U/dt &= +k^+ n_N - k^- n_U \end{cases} \tag{2.11}$$

When the kinetic rates are constant, the solution for $t \longrightarrow \infty$ has an exponential decay, with a time constant equal to $|k^+ + k^-|$, attaining the *equilibrium concentration* determined by the condition $dn_N/dt = dn_U/dt = 0$:

$$k^+ n_N = k^- n_U \tag{2.12}$$

This is the *detailed balance condition* meaning that the forward flux from the state N to U is equal to the inverse flux from U to N.

Arrhenius supposed for the unidirectional forward process that the reactants transforming into products are in an activated state $\bar{N}$ at the equilibrium with the rest of reactants. In this condition, using the Van't Hoff equation he derived a temperature-dependent relation for the forward kinetic rate:

$$k^+ = k_0 \exp(-\frac{\Delta E}{k_B T}) \tag{2.13}$$

where k_0 is a pre-exponential factor and ΔE is the temperature-independent energy difference between the active state and the fundamental state.

2.2.1 *Kramers theory*

Further works tried to give a more formal justification the Arrhenius description, in connection with the theory of particle diffusion through a potential barrier. If the system (let us think of an ensemble of particles in a potential minimum) is initially in a thermalized state N, and each particle has to overcome an energy barrier $E_B \gg k_B T$, the system very slowly relaxes into the new minimum U of the potential, since the jump probability for each particle is exponentially small. In this case a description in terms of *rare events* makes sense; otherwise,

the particle would be free to jump back and forward through the barrier (i.e., the event is "rare" with respect to the picosecond time-scale of molecular thermal motions, although it may not be rare at all from the macroscopic point of view). In this case we can define a stationary flux from the initial state to the limiting state B that defines the barrier.

In his 1940 paper *Brownian motion in a field of force and the diffusion model of chemical reactions* [82], A. H. Kramers used the Fokker-Planck equation describing the Brownian motion in phase space for a generalized coordinate x. This unique coordinate describes the transition process, while all the others degrees of freedom of the system represent the thermal bath at temperature (T). In this way he could define the stationary flux j, representing the rate of escape from the potential barrier. To do so, let us consider the coordinate dynamics governed by a Langevin equation:

$$M\ddot{x} = -U'(x) - \gamma M\dot{x} + \eta(t) \tag{2.14}$$

where M is the *effective mass* of the particle, $V(x)$ is the potential energy function, γ is a friction coefficient defining the interaction of the particle with the thermal bath, and $\eta(t)$ a Gaussian white noise, with the properties:

- $\langle \eta(t) \rangle = 0$
- $\langle \eta(t) \cdot \eta(s) \rangle = 2M\gamma k_B T \cdot \delta(s-t)$

A consequence of this assumption is that the system dynamics is Markovian. This can be a restriction if we are dealing with generalized coordinates for which memory effects could occur, but it could be generally be assumed in the limit where the other degrees of freedom behave like a proper thermal bath.

It is possible to rewrite this equation in terms of the corresponding probability distribution function $p(x, v, t)$, describing the probability of finding a particle at the coordinate x at time t with velocity v. Then, to construct the stationary flux the following assumptions are made:

- a source supplies the N-state potential well with particles at energy just below the barrier, located at the distance $x = x_B$, such that the population in the initial state is constant, and the process of overcoming the barrier is due only to thermal fluctuations and not to an external flux of "hot" particles;
- the rate of escape from the well is slow compared with the thermal kinetics inside the initial state, i.e. the particle density in the N state is:

  $$\rho(x, v) = Z^{-1} \exp\left[-\tfrac{1}{k_B T}\left(\tfrac{M}{2}v^2 + U(x)\right)\right], \text{ in } x < x_B \ ;$$

- the particles that eventually overcome the barrier are immediately removed, so that no reverse flux from the U to the N state is observed:

$$\rho(x, v) \approx 0, \text{ in } x \approx x_U \quad .$$

The reactive flux is then defined as the flux of particles going through the barrier, $j = \int v\rho(x_B, v)dv$. The population of particles in the N state is $n_N = \int_{x<x_B} dx \int dv\rho(x, v)$, if the native state N is assumed to be in the region $x < x_B$. For a spatial diffusion-limited flow, the kinetic rate is then equal to:

$$k^+ = \frac{j}{n_N} = \int_{-\infty}^{x_+} dy \frac{M\gamma}{k_B T} \exp[\beta U(y)] \int_{-\infty}^{y} dz \exp[-\beta U(z)] \quad (2.15)$$

In the saddle-point (Taylor) approximation, the above equation takes the form:

$$k^+ = \frac{\lambda_+}{\omega_B} \frac{\omega_N}{4\pi} \exp[-\beta E_B] \quad (2.16)$$

In the previous expression, $\lambda_+ = -\gamma/2 + \sqrt{\omega_B^2 + (\gamma/2)^2} \leqslant \omega_B$, and $E_B = U(x_B) - U(x_N)$ is the energy barrier to overcome. The ω factors are the oscillator frequencies at the coordinates x_N and x_B, where the shape of the binding potential is locally approximated by two harmonic oscillators with curvature $k = M\omega^2$.

In the limit of *strong friction*, that is $\ddot{x} << \gamma\dot{x}$, the well-known overdamped regime formula is obtained:

$$k^+ = \frac{\omega_N \omega_B}{4\gamma\pi} \exp[-\beta E_B] \quad (2.17)$$

From the shape of the potential energy along the reaction coordinate, one can then determine the evolution of the population of each state (it would be concentrations of reactants and products in a chemical context). In the Fig.13 we show a (purely mathematical) example of how the equilibrium population distributions are related to the kinetics of the reaction.

Unfortunately, due to the exponential dependence on the energy barrier, even a small difference in energy E_B between the two molecular states could require a huge amount of time, before observing a transition from N to U. This is not always a limitation for a macroscopic sample, where the number of particles is of the order of 10^{23}, but definitely creates a problem in single-molecule experiments, where just one, or a few molecules at a time are considered. Hence, the experimental time of observation must be long enough to observe the transition(s). For example, in a DNA hairpin of 10-20 bp at physiological conditions, similar to the ones we used in our experiments, it could require years to observe a single transition. It is then necessary to find a method to boost the system kinetics, and force the system to overcome the barrier.

2.2.2 *Bell-Evans theory*

Starting from the results obtained by Kramers, in a seminal paper appeared in 1997 Evans and Ritchie [50] considered the effect on the kinetic rates due to an external force applied on the system. The first

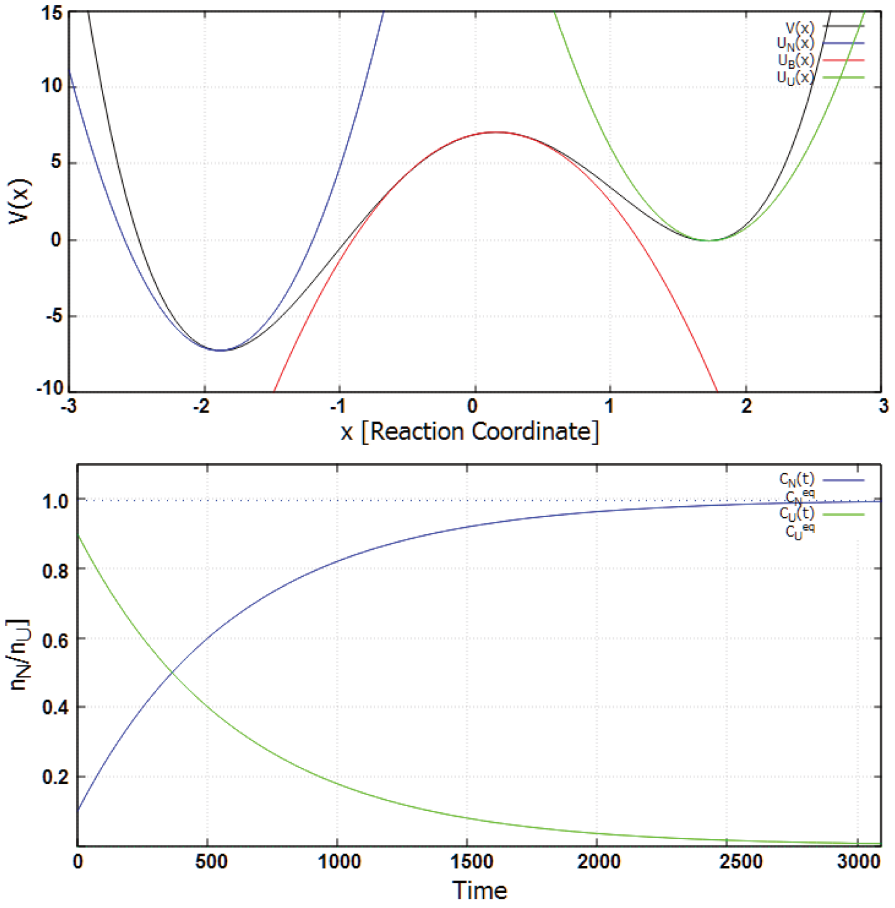


Figure 13: Demonstration of the reaction rate theory for an asymmetric double-well potential. This is a purely mathematical example, so no physical units are used in the plot. The potential energy has been approximated with its Taylor expansion around the extremants. On the negative part of the plot the state N (blu), on the positive side of the reaction coordinate the U state (green) and between them, close to the zero, the potential barrier (red). Solving the differential equation (2.11) for the the two concentrations (bottom) with initial values $n_N = 0.1$ and $n_U = 0.9$ we obtain the time-evolution of the systems, that moves from a prevalent U concentration to a prevalent N.

to introduce this idea was G. Bell about 15 years before[1], in a much cited Science paper [17] where he found an exponential behaviour for single-bond lifetime in cell membranes held together by point ligands:

$$k^+ = \frac{1}{t_{off}} \exp(f/f_\beta) \tag{2.18}$$

where t_{off} is the average time of bond breaking under zero applied force, and f_β is a thermal coefficient with the dimension of a force.

Evans and Ritchie considered the process of bond formation in a fluid as a system out of equilibrium trapped in a local minimum. As prescribed from the Kramers theory, the process of escape from the minimum could be described as a constant diffusive flux of thermalized states. When an external force is introduced, it acts like a "path selector" and it becomes possible to describe the kinetic rate of the escape process by a generalized scalar coordinate:

$$k^+ = \frac{\omega_N \omega_B}{4\gamma\pi} \exp(-\beta E_B(f)) \tag{2.19}$$

The external force distorts the entire potential energy surface from the initial $V(x)$ to some $V'(x, f)$, as shown in Figure 14 for the simple case of an asymmetric double-well potential. The most important result is to affect the probability to reach the top of the barrier. In the simplest case, the variation in energy barrier height E_B^0 is approximated as a linear factor proportional to the applied force:

$$E_B(f) = E_B^0 - x_{N-B} \cdot f \tag{2.20}$$

x_{N-B} being the projection of the distance between the bound state x_N and the energy barrier x_B along the axis parallel to the vector of the force. This approximation is verified for a sharp barrier if the curvature of the potential at the transition state is not affected by the applied force.

The system kinetic rate can be expressed as a function of the external force applied:

$$k^+(f) = \frac{\omega_N \omega_B}{4\gamma\pi} \exp[-\beta(E_B^0 - x_{N-B} \cdot f)] \tag{2.21}$$

Using the external force the system can therefore be pushed into a particular state, and by varying the modulus and direction of the force applied it is possible to explore the free-energy landscape of the system. The analogy between the potential well with a barrier, and a chemical bond that can be broken or reformed by scanning up and

1 A former PhD student of Hans Bethe, George Irving Bell gained international recognition as a theoretical nuclear physicist, contributing to the solution of the neutron transport problem in nuclear reactors, and authoring with Samuel Glasstone the most famous book *Nuclear Reactor theory* (1970); he turned to biology around 1960, publishing many relevant papers and two more books, and funding firstly the Los Alamos TBB group (Theoretical biology and biophysics), then in 1988 the Center for Human Genome Studies in Bethesda.

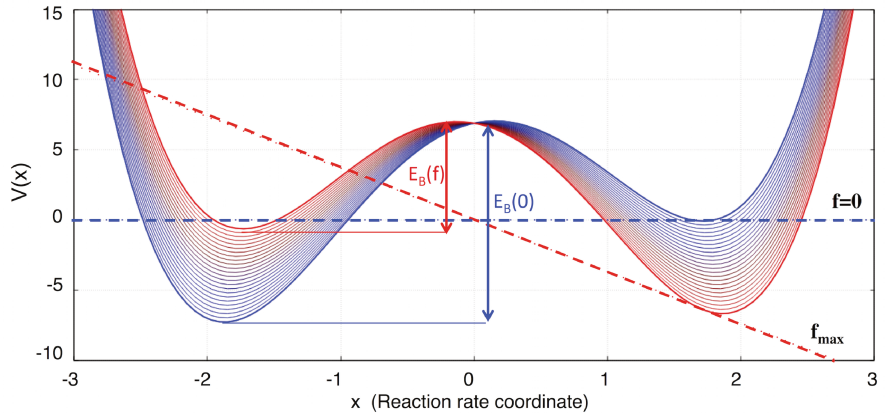


Figure 14: Effect of an external force on an asymmetric double-well potential. Using the same potential of Fig.13 (top) represented in blue, an external force (dashed) is applied on the system, until a maximum value f_{max} is reached. The effect of the external field is to "rotate" the initial potential, proportionally to the applied force (red curve). The positions of states N, U and the barrier B are little modified, the main effect is to change the energy barrier between the minima and the top of the barrier, $E_B(0)$ and $E_B(f)$.

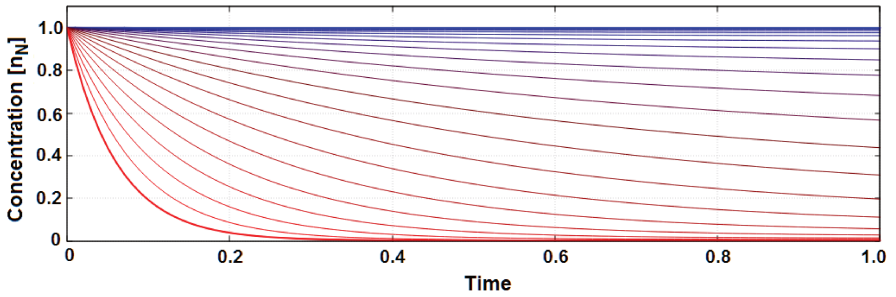


Figure 15: Reactant concentration for an asymmetric double-well potential in presence of a external force. As in Fig.14, the blue curve represents the concentration of the N state as a function of time at zero force, for an initial concentration $n_N = 1$, $n_U = 0$. Moving from blue to red, the force is increased and the equilibrium concentration increasingly moves to the U state.

down the force between 0 and f_{max}, means that one can obtain information on regions of the energy landscape that may be not accessible in the experimental time-scale. As an example, the concentration time-evolution for the potential in Fig.15 are modified, as the equilibrium concentrations are moved.

In the next Section we will apply the concepts developed by Bell, Evans and Ritchie, in the context of bond rupture and formation during single-molecule force spectroscopy experiments. To this purpose [2], it is useful to consider also the reverse reaction:

$$k^{-}(f) = \frac{\omega_U \omega_B}{4\gamma\pi} \exp[-\beta(E_B^0 - \Delta G_{NU} - x_{B-U} \cdot f)] \tag{2.22}$$

and contract the pre-exponential factors in Eq.(2.21) and 2.22) into a factor k_m^+, k_m^-, assumed constant in the neighborhood of f. The forward and backward reaction rates become:

$$k^+(f) = k_m^+ \exp[-\beta(E_B^0 - x_{N-B} \cdot f)] \tag{2.23}$$

$$k^-(f) = k_m^- \exp[-\beta(E_B^0 - \Delta G_{NU} - x_{B-U} \cdot f)] \tag{2.24}$$

By taking the logarithm of the ratio between the two:

$$\begin{aligned} \ln\left(\frac{k^+(f)}{k^-(f)}\right) &= \ln\left(\frac{k_m^+}{k_m^-}\right) - \beta\left[\Delta G_{NU} - (x_{N-B} + x_{B-U}) \cdot f\right] = \\ &= \ln\left(\frac{\omega_N}{\omega_U}\right) - \beta\left[\Delta G_{NU} - (x_{N-B} + x_{B-U}) \cdot f\right] \end{aligned} \tag{2.25}$$

Because of the detailed balance condition, Eq.(2.12), this quantity is also equal to $\ln(n_U^{eq}/n_N^{eq})$ at equilibrium.

Finally, the variation of the potential barrier induced by the external force is:

$$\Delta E_B(f) = \Delta G_{NU} - (x_N - x_U) \cdot f \tag{2.26}$$

2.2.3 *Recovering free-energy differences from single-molecule hopping experiments*

In a hopping experiment, the molecule is held in the optical trap at a foce value such that the probabilities of forward and backward jump between two states are approximately equal. Therefore, the "hopping" event is observed with a relatively high frequency. In the process of bond formation and breaking, we start from the general definition of Gibbs free-energy $G(p,T) = U + pV - TS$, and consider the differential relation:

$$dG = Vdp - SdT + \sum_{i=N,U} \mu_i dn_i - xdf \tag{2.27}$$

where n_i is the number of particles occupying a given state (for the sake of simplicity we consider only the two states N and U), μ_i is the corresponding chemical potential, x is the generalized coordinate of the reaction (the one once used to describe the reaction rate), and f is the force applied along the reaction path.

In an open system where temperature and pressure are kept constant, the relation becomes:

$$dG = \sum_{i=N,U} \mu_i dn_i - xdf \tag{2.28}$$

If the system does not change its state along the transformation, the finite variation of free-energy is equal to the reversible work done on the system:

$$\Delta G = -\int xdf = \Delta W^{Rev} \tag{2.29}$$

while for a transformation in which the external force is constant, the variation of free-energy is:

$$\Delta G = \mu_N \Delta n_N + \mu_U \Delta n_U \tag{2.30}$$

which, if we consider only one bond in a single molecule,

$$\Delta G_{NU} = \mu_U - \mu_N \tag{2.31}$$

is just the free-energy of bond formation.

2.2.4 *Bond-rupture rate in pulling experiments*

Let us now consider a single-molecule experiment in which a molecular system, maintained in the initial state N at a given force f_N, is pulled towards a different final state U by the external force at constant pulling rate, $R = df/dt = const$. In this case, the kinetic equation (2.11) as a function of the applied force can be rewritten as:

$$-k^+ n_N + k^- n_U = \frac{dn_N}{dt}\frac{df}{dt} = \frac{dn_N}{df}\frac{df}{dt} = R\frac{dn_N}{df} \tag{2.32}$$

If the occupation probability of the final state is kept fairly small, $k^- n_U \sim 0$, the solution is:

$$n_N(f) \simeq n_N(0) \exp\left[-\frac{1}{R}\int_{f_1}^{f} k^+(f')df'\right] \tag{2.33}$$

The probability of observing the bond rupture at any force $f > f_N$ is given by the force-dependent variation of the relative occupation of the initial state:

$$p_N(f) = \frac{dn_N(f)}{df} = -\frac{k^+(f)}{R} n_N(f) \tag{2.34}$$

In "pulling" optical-tweezer experiments it is possible to directly observe the value of force at which the system experiences a *first transition* to the unbound state U, starting from the initial state N. By cyclically repeating the force-displacement trajectory many times during the same experiment, a probability distribution of the values of rupture force can be constructed. Since we are considering just the first bond rupture, and the system is initially stable in the N state, the above stated condition $k^- n_U \sim 0$ is verified. Therefore, we could extract the kinetic rate from the probability rupture force distribution, and interpret such values with the Bell-Evans theory.

Combining the result from the Eqs.(2.23)-(2.24), and Eq.(2.34):

$$k^+(f) = -R\frac{p_N(f)}{n_N(f)} = k_m^+ \exp[\beta x_{N-B} \cdot f] \tag{2.35}$$

where the constant k_m^+ represents the kinetic rate at zero force, and x_{N-B} the projection along the force direction of the distance between the bound state and the barrier position (i.e., the transition state).

The same procedure can be repeated for the process of bond formation, during a pulling experiment with the system is initially kept in the unbound state U, and where the force is constantly decreased. The force at which the bond is formed for the first time can be recorded on the experimental force-displacement trajectory, and a similar relation between the *first binding force* and the kinetic rate can be obtained (in the following we will refer to both values as "rupture force" to avoid excessive notation).

By comparing the two kinetic rates (actually, their exponential factors in Eq.(2.35)) for the separate experiments of bond rupture and bond formation, the *coexistence force* value can be extracted; moreover, we obtain information about the potential barrier length (that is, the $N - B$ and $B - U$ distances), and the free-energy of transition at zero force. In the next Chapter we will use these considerations to extract information on the DNA hairpin free-energy landscape, and to determine the variation of free-energy due to the presence of a mismatch defect in the hairpin sequence.

2.3 THE DNA HAIRPIN

The *stem-loop* configuration, or "hairpin", is a DNA sequence constructed so that intramolecular pairing of self-complementary regions occurs (Figure 16). Such a configuration is a key building block of many RNA secondary structures, but is also present in DNA especially in poly-AT regions, where transition from the usual double-stranded helix to a hairpin could spontaneously occur. Nucleic acids hairpins consist of a stable segment of ssDNA (or ssRNA), including two self-complementary strands at the $5'$ and $3'$ ends, which fold into a double helix (hereafter referred to as the *stem*), and a variable-length central region of nucleotides that upon folding form a *loop*.

The stability of the hairpin depends on the binding energy between the bases that form the stem and the unpaired loop termination. As a consequence, this is strongly sequence-dependent, and defects in the structure like methylation or insertion of a mismatch may change the stability of the entire sequence. A simple way to estimate the hairpin free-energy of formation is the *nearest-neighbor model* [30, 130], an empirical model that assigns a contribution to the total energy depending on the composition of each base-pair and its immediate neighbor bases, plus penalty terms for the unpaired loop termination and initial base-pair. Tables of energy values have been experimentally established for each possible combination of base-pairs and neighbors, traditionally by means of DNA melting experiments (calorimetry, [118, 130, 131]) and in recent times also by the more advanced technology of DNA microarrays [66].

Their sequence-dependent stability, and the capability to easily force the repeated unfolding/refolding of the hairpin into/from a single stranded DNA molecule, has made these molecular constructs one of the emblematic system for single-molecule dynamic force spec-

troscopy. During the experiments, the force is applied at the two ends of the construct and, by measuring the force-extension relation, information on its free-energy landscape are obtained.

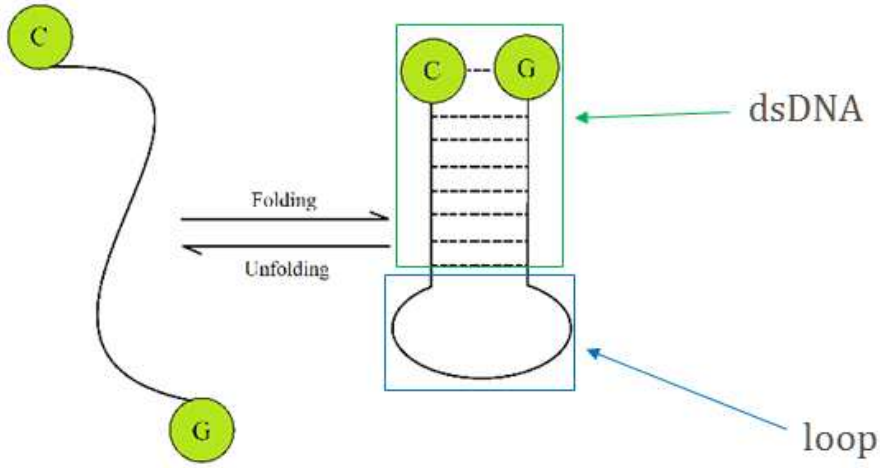


Figure 16: Schematic representation of a DNA hairpin: on the left, the DNA strand is unfolded and behaves like a ssDNA polymer chain; on the right, the self-complementary bases have folded and formed the hairpin structure, a dsDNA region terminated with a loop of unpaired bases.

In all this kind of single-molecule experiments the molecule of interest cannot be directly attached to the force actuators (the AFM cantilever tip, the micro beads in magnetic and optical tweezers), therefore it is always necessary to use linker molecules as intermediates, or spacers. For example in the experiments that will be described in the following Sections 3.5 and 3.6, the hairpin was sandwiched between two dsDNA handles with identical sequence, to simplify the data analysis. The length of the handles could vary depending on the necessity of the experiment, but is usually kept quite short to reduce the noise on the small hairpin (in our case, 29-basepairs) [53].

The intermediate dsDNA handles are the ones to be directly attached to the surface of the microbeads. To avoid non-specific adhesion between the two beads and between the beads and the molecular construct in a non-terminal site, the binding of the construct (2 handles + 1 hairpin) to the beads is achieved via specific biochemical reactions (antibody-antigen bonds). In our experiments two specific groups are inserted at the free end of each of the two handles (Figure 17):

- the 5'-extremity of the DNA sequence is functionalized with a biotin, a species often used in molecular biology experiments thanks to its strong non-covalent interaction with streptavidin, which is resistant to temperature, pH changes and other denaturation factors; streptadivin-coated beads will be inserted directly into the microfluidic chamber;
- the 3'-extremity is functionalized with a digoxigenin, which has the role to selectively bind to the surface of beads functionalized

with the antidigoxigenin enzyme in a preparatory reaction, usually carried out just before the experiments (the procedure takes about 20 min); the molecular construct attached to this bead by the 3′ end is then inserted in the chamber, and immobilized by a suction micropipette.

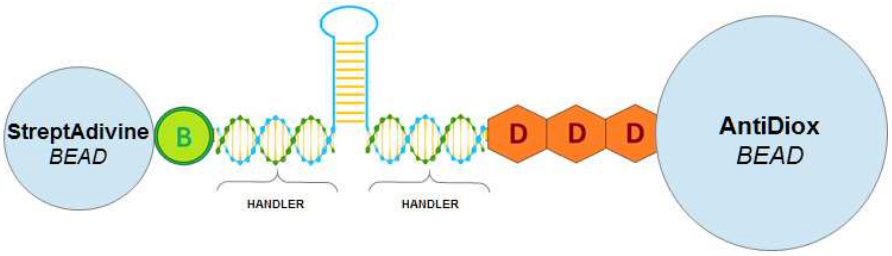


Figure 17: Schematic representation of the dsDNA+hairpin construct used in the experiments. The two microbeads are functionalized with streptavidine and anti-dioxigenin antigens, to attach to the antibody biotin (B) and dioxigenin (D), respectively placed on the two ends of the dsDNA handles. The larger microbead (right in the Figure) is the one to be captured in the optical trap, whereas the smaller one is held at a fixed position by a suction micropipette.

Using these two specific bonding reactions, it is possible to trap the DNA between the two coated beads, avoiding that both terminations could stick to the same bead. Evidently, the relation between force and extension obtained from the instrument needs to be filtered, to take into account the effects of the handles and separate them from the hairpin properties. This can be done by a theoretical model, constructed with the elements described in the previous two Sections.

2.4 THEORETICAL MODEL FOR THE MOLECULAR SYSTEM

One important principle underlying the mechanism of optical trapping, which will be described in full details in the Section 3.2 below, is that when a force is applied on the trapped particle, it induces a displacement of the bead inside the trap (see Figure 18). The introduction of the spacer handles adds another contribution to the extension measured by the instrument, which must be properly accounted for. The scheme in Fig.18 suggests the ingredients for a model capable of describing the relation between the measured force and extension, and the hairpin intrinsic elastic properties.

Said λ the total extension measured by the optical tweezer under a force f, this quantity can be ideally decomposed into the sum of three contributions:

$$\lambda(f) = x_{bead}(f) + x_{handles}(f) + x_{DNA}(f) \tag{2.36}$$

The terms of the equation are explicited as follows:

- x_{bead} is the displacement of the moving bead inside the optical trap. The optical trap generates a potential that is well approx-

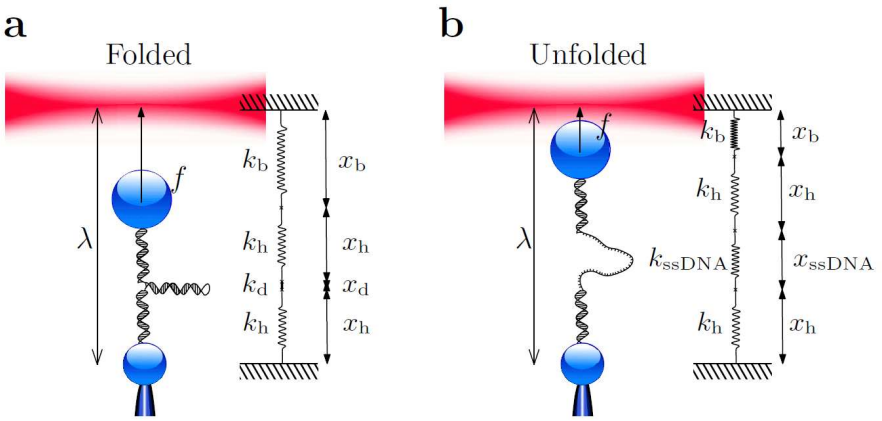


Figure 18: Schematic illustration of the contributions to the total trap-pipette distance in single-molecule pulling experiments performed on a DNA hairpin with the mini tweezer. The measured extension is the sum of the displacement of the mobile bead in the optical trap, the elongation of the handles, and the extension of the hairpin. This last term depends on the configuration of the hairpin: in the folded state (a), it contributes only with the orientation of the folded DNA diameter; in the unfolded state (b) the stretching of the hairpin under opening contributes to the total elongation. Image courtesy of Anna Alemany.

imated by an harmonic spring, so the displacement is linearly related to the force applied: $\Delta x_{bead} = \Delta f / k_{trap}$.

- $x_{handles}$ is the contribution due to the two handles. Since both handles are a dsDNA chain, it is possible to describe their elastic response by an adequate polymer model as the worm-like chain.

- x_{DNA} is the contribution due to the hairpin, which depends on the folding state of the hairpin:
  - In the *folded state*, the end-to-end distance (that is, the distance between the green-circled G and C bases in Fig. 16) equals the dsDNA diameter x_d; its contribution to the total extension is just given by the projection of x_d along the direction of the force. During the experiments, the force is chosen to be aligned with the $\hat{y}$ axis, while the contribution along the perpendicular $\hat{x}, \hat{z}$ directions reduce to a Gaussian noise with a zero average. We can use the FJC model for a single-monomer chain, Eq.(2.4) above, to calculate the expectation value of the work done to orient a segment of fixed lenght along the applied force vector.
  - When the hairpin begins to unfold, the contribution of the partially unfolded fraction (the base-pairs next to the terminal ends) must be taken into account. Each unfolded segment of n base-pairs can be described as a two independent polymer chains of nucleotides (neglecting the chemical differences), for which the WLC model is applied.

– Finally, when the last bases of the stem get opened, the entire molecule contributes as one single-stranded DNA chain composed by the total number of bases in the hairpin (stem + loop).

Even if we are dealing with a system composed by just one molecule, a statistical mechanics description of the system behaviour is justified by the fact that the polymer molecule is forced to explore the configuration space by the contact with the surrounding solvent. The impact between the solvent molecules and the chain forces the latter to explore different configurations, as if being perturbed by a white noise. Because the solvent mass is so much larger than the molecule mass, the solvent behaves like a thermal bath for the molecule, at constant temperature if its temperature is maintained fixed. The polymer chain can then described by an ensemble with a fixed number of chain elements, constant temperature, and an external force that pulls the system along a direction. A similar polymer model decomposition could be used for experiments on RNA, proteins or other biomolecules in general. What would change is the specific model use to describe each particular contribution. In the next Chapter we will precisely identify the explicit contributions used to model our hairpins in optical tweezer experiments.

3

DYNAMIC FORCE SPECTROSCOPY

3.1 SINGLE-MOLECULE TECHNIQUES

Substantial efforts have been devoted in recent years to experimentally characterize the free-energy landscape of biomolecule folding (see e.g. [44, 126, 152] and references therein). Dynamic force spectroscopy experiments are well-suited to study the folding/unfolding transitions of one molecule at a time with high spatial and temporal resolution, as a function of the molecular deformation [20, 22, 141]. By applying a controlled force to the ends of a molecule one can observe the mechanical unfolding and folding transition, eventually unraveling the presence of intermediate and misfolded states. Additionally, from the experimentally measured force-dependent unfolding and folding kinetic rates it is possible to characterize the position of the transition states, and the height of the corresponding kinetic barriers [126]. Notably, force spectroscopy by optical tweezers has been successfully applied also to measure very small molecules, such as short DNA hairpin sequences, thereby allowing to detect elastic properties of ssDNA [4], and transition states during their repeated folding/unfolding [3].

Giving access to the properties of one single molecule at a time, force spectroscopy should be well suited also to study the impact of structural defects in the molecular edifice, by comparing the force-extension relations and free-energy landscape of the pristine vs. the defective molecule. Until now, however, this possibility has been very little explored. A few works have studied the impact of point mutations or topological alterations in proteins [8, 32, 136, 166], but almost nothing exists in literature for nucleic acids. Only in a recently published study [74], optical-tweezer force induced unfolding was used to characterize four different mismatch defects in DNA, by positioning pairs of identical defects at close distance in 29-bp DNA hairpins with a 6-bases loop.

In the experimental part of this thesis, we used the same technique of force spectroscopy with optical tweezers, to demonstrate the capability to detect the excess free energy and transition-state features of individual mismatches in DNA, at the scale of the single molecule. The mismatch defect was chosen because it is structurally simpler, compared to e.g. crosslink or strand breaks; moreover, strand breaks (which will be at the focus of the second part of this thesis) are in principle impossible to study, since they imply an already broken or very weakly bound molecule. Mismatches can, instead, be easily constructed by design, at the moment of ordering a specific sequence of DNA from the supplier. For this study, we included properly constructed single G-A or G-T mismatches in very short DNA hairpins, of 10 to 20 bp length, tethered by two fragments of dsDNA

also of very short length (29 bp each). Monitoring the folding/unfolding transition of the hairpin under a constant, or a linearly-variable external force, allows to extract information about the binding free-energy, coexistence force (at which the folded/unfolded states have equal occupation probability), and internal dynamics of the pristine and defective molecule. The results to be discussed in the second half of this Chapter will demonstrate the ability of this experimental method to clearly detect the presence of the mismatch defects in the DNA sequence; defect energetics and dynamics can be qualitatively characterized, and also quantitatively within the lower limits of the experimental resolution.

3.2 A FOCUS ON OPTICAL TRAPS

Arthur Ashkin observed in the 1970 [9, 10] that by using a highly focused laser beam it is possible to trap and move microsized particles suspended in a fluid. Since the original statement, this technique has been enormously developed and adapted to use in many different fields, ranging from from condensed-matter physics (e.g. to cool atoms in Bose-Einstein condensation), to biology (single-molecule and single-cell manipulation). After the Nobel Prize to Steven Chu, Claude Cohen-Tannoudji and William Phillips in 1997 for "development of methods to cool and trap atoms with laser light", and a second Nobel Prize with a strong implication of optical traps arrived in 2001, to Eric Cornell, Walter Ketterle and Carl Wieman, optical traps still continue to find new applications, which ended up in the third, 2018 Nobel prize just a few weeks ago to Ashkin, exactly "for the optical tweezers and their application to biological systems".

Different theories describe the interaction between matter and light, ranging from the simple model of Mie scattering to the fundamental equations of Quantum Electrodynamics, and depending on the characteristics of the light and particles involved it may be appropriate to formulate the description in a different framework. Maxwell equations describe the electromagnetic field and, as a general consequence of them, it can be stated that light carries momentum. During the interaction between matter and light there may be some transfer of momentum, according to the law of conservation and the corresponding kinematics. In a simple semi-classical model, a particle of mass m interacts with an incoming photon of momentum $\vec{p}_{\lambda_{in}}$ that changes to $\vec{p}_{\lambda_{out}}$ after the interaction, while the particle changes its velocity by $\Delta\vec{v}$. For the conservation of momentum:

$$\Delta\vec{v} = \frac{1}{m}\left(\vec{p}_{\lambda_{in}} - \vec{p}_{\lambda_{out}}\right) \tag{3.1}$$

In the most basic case of interaction, the photon is simply absorbed by the particle and $\vec{p}_{\lambda_{out}} = 0$:

$$\Delta\vec{v} = \frac{h}{m\lambda}\hat{k} \tag{3.2}$$

where the unit vector $\hat{k}$ is the direction of propagation of the photon, h the Planck constant and the relation between photon momen-

tum and wavelength λ is the standard quantum-mechanical relation $p_\lambda = h/\lambda$. Therefore, the velocity change is related to the photon wavelength λ. If the process of photon absorption is repeated in time at a rate $dN(t)/dt$, the particle will accumulate a force directed along the photon wavevector:

$$m\frac{d\vec{v}}{dt} = \frac{h}{\lambda}\frac{dN(t)}{dt}\hat{k} = \vec{f} \tag{3.3}$$

The number of photons absorbed depends on the intensity of the radiation field, I_0, the energy of the single photon E_λ, the interaction surface offered by the object to the incoming beam, S, and a coefficient of absorbance ζ depending on the particle material:

$$\frac{dN(t)}{dt} = \frac{S\zeta}{E_\lambda}I_0 \tag{3.4}$$

For a monochromatic source ($E_\lambda = hc/\lambda$), combining the two equations (3.3) and (3.4):

$$\frac{d\vec{v}}{dt} = \frac{S\zeta}{m}\frac{I_0}{c}\hat{k} = \sigma\frac{I_0}{c}\hat{k} \tag{3.5}$$

The previous equation defines, for a homogenous, spherical particle of radius r and density ρ, the *light absorption cross section*:

$$\sigma = \frac{3}{4\pi}\frac{\zeta}{\rho r^3} \tag{3.6}$$

The two equations above show that the acceleration imparted by light absorption is imperceptible for macroscopic objects, because of the size and mass at the denominator, i.e. the body inertia and thermal noise completely overwhelm the light source intensity. However, this is not the case at the microscopic scale. In biophysics experiments the particle diameter is typically ~1-2 μm, a size that ensures a reasonable light absorption and momentum transfer.

Obviously, this simple model is not always applicable to the interaction between light and matter, for example light could be just scattered in a different direction, or traverse the particle and change its momentum (refraction), or could be composed by a distribution of wavelengths each one interacting differently with the particle. For the range of energies, time-scales and length-scales typically used in biophysics experiments, the interaction between matter and electromagnetic waves are correctly described by the classical electromagnetic theory. Gustav Mie in 1908 was the first to obtain a rigorous solution for the diffraction of a plane wave by a homogeneous sphere. Further developments originated three theoretical approaches to the subject, depending on the ratio between the light wavelength and the size of the particle:

- for $r \ll \lambda$, the Rayleigh regime: the particle can be treated as a small spherical dipole with a uniform electromagnetic field.
- for $r \approx \lambda$, an intermediate regime arises, which can be described by the *generalized Lorenz-Mie theory* (see e.g. ref.[59]);

- for $r \gg \lambda$, the Mie regime (or *ray-optics*): the light beam is decomposed into individual rays characterized by their intensity and direction, following the laws of geometrical optics;

The generalized Lorenz-Mie theory is the model that better describes the physics of optical trapping in the regime of interest for biophysical experiments, since particle sizes are of the order of a few μm and the commonly used laser wavelengths are in the range $\sim 800 - 1200$ nm. This theoretical approach, however, requires a complete solution of Maxwell's equations with the appropriate boundary conditions. Therefore, to gain at least a qualitative description of the principle of optical trapping, we will discuss below the two limiting cases of the Mie, and Rayleigh regime.

3.2.1 *Rayleigh regime*

The Rayleigh regime considers the limit where the particle is very small compared to the wavelength. In this condition, the perturbation produced by the particle on the wavefront are negligible, and the particle can be described as a point dipole. The force acting on the particle can be divided in two components [114]:

- *Scattering force* ($\vec{f}_{scattering}$), due to the absorption and re-radiation of the light radiation pressure on the particle. The light absorbed from the material is re-emitted isotropically by the atoms (or molecules), so the difference between the photons absorbed and emitted creates a net force pushing the particle along the beam propagation axis ($\hat{k}$):

$$\vec{f}_{scattering} = n_m \frac{\sigma}{c} I_0 \hat{k} \tag{3.7}$$

  where n_m is the refractive index of the surrounding media, and σ is the particle cross section, which for a small homogenous spherical dipole of radius r is equal to:

$$\sigma_{sph} = \frac{128\pi^5 r^6}{3\lambda^4} \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \tag{3.8}$$

  In the latter equation $n = n_{sph}/n_m$ is the ratio between the refractive index of the particle and that of the medium.

- *Gradient force* ($\vec{f}_{gradient}$), due to the force acting on the particle-induced dipole $\vec{d} = \alpha \vec{E}$ (with α the *polarizability*):

$$\vec{f}_{gradient} = \frac{2\pi\alpha}{c n_m^2} \vec{\nabla} I_0 \tag{3.9}$$

  Its dependence on the intensity gradient allows to shape the trap by using optical components, such as a lens. The polarizability describes the type of equilibrium around the intensity maximum:

- $\alpha > 0$ implies the equilibrium is *stable* so the particle tends to return in the position of maximum intensity;
- $\alpha < 0$ implies the equilibrium is *unstable* so the particle is repelled from the maximum intensity.

For a homogenous sphere the polarizability is: $\alpha = n_m^2 r^3 \frac{n^2-1}{n^2+2}$, therefore if the ratio $n > 1$ the force attracts the particle towards the maximum. Such a difference between the refractive index of the particle and the medium is actually the key for the principle of optical trapping.

Using a highly focussed beam it is possible to shape the gradient force so as to contrast the effect of the scattering force, thereby preventing the particle to be pushed away in the direction of the light propagation. This can be achieved by using a high numerical aperture (NA) objective lens (see Figure 86 in the Appendix A). In modern instruments it is possible to combine the effects of high-NA objective with the contribution of a counter-propagating laser beam, to cancel out the scattering force and increase the trapping force along the propagation axis.

3.2.2 *Ray-optics regime*

In this approximation, the light wavelength is much smaller than the typical dimensions of the object, and the undulatory behaviour is therefore negligible. The light beam can be decomposed into rays that propagate in straight line and are described by their momentum, direction and intensity. The rays obey the laws of geometrical optics,

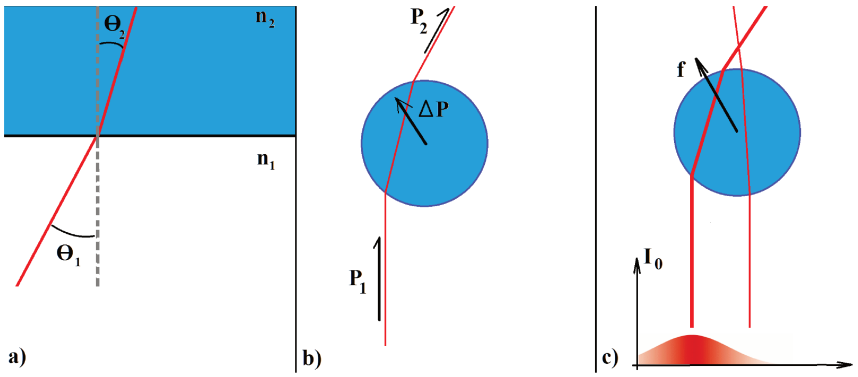


Figure 19: Geometrical optics. (a) Passing through the interface between two materials, the direction of a ray change accordingly to the Snell's law (Eq.3.10) and the refracted angle θ_2 depend on the incident angle θ_1 ad the rate between the refraction index of the two materials. (b) Change in ray direction passing through a spherical particle of refractive index greater than the surrounding media. (c) A Gaussian beam due to the different intensities of the rays produce an effective force on the particle that tend to align the center of the sphere with the maximum of the beam (restoring force).

and modify their path when passing through a surface according to Snell's law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{3.10}$$

where n_1,n_2 are the refraction index of the two materials (e.g., water and plastic microsphere), and θ_1,θ_2 are the angles formed between the surface and the ray direction (see Figure 19(a)).

For a rigid spherical particle made from a homogeneous material with refraction index greater than that of the surrounding medium, the rays are deflected according to the scheme in Figure 19(b). The variation of the ray direction corresponds to a transfer of momentum, from the light beam to the particle. Using different light detectors is it possible to measure the variation on the beam and deduce the force applied on the particle.

So, if we consider the intensity profile of a Gaussian beam, the different intensities of light rays produce a net force on the particle. The orthogonal component of this force (with respect to the beam propagation) is called by analogy with the Raylegh regime, gradient force, and the parallel component is analogous to the scattering force. The gradient force is null when the particle is in the region of maximum intensity of the beam, while it is non-zero and tends to restore the particle to the maximum when its position gets displaced (Fig.19c).

On the other hand, if the beam is focused, for example by introducing a convergent lens along the optical path, the intensity has a maximum localized in the focal point. This intensity profile can be shaped in such a way to trap the particle (see the example given in Figure 86 in the Appendix A.1). The displacement of the particle from the rest position inside the trap can be detected by measuring the refracted light (see Figure20).

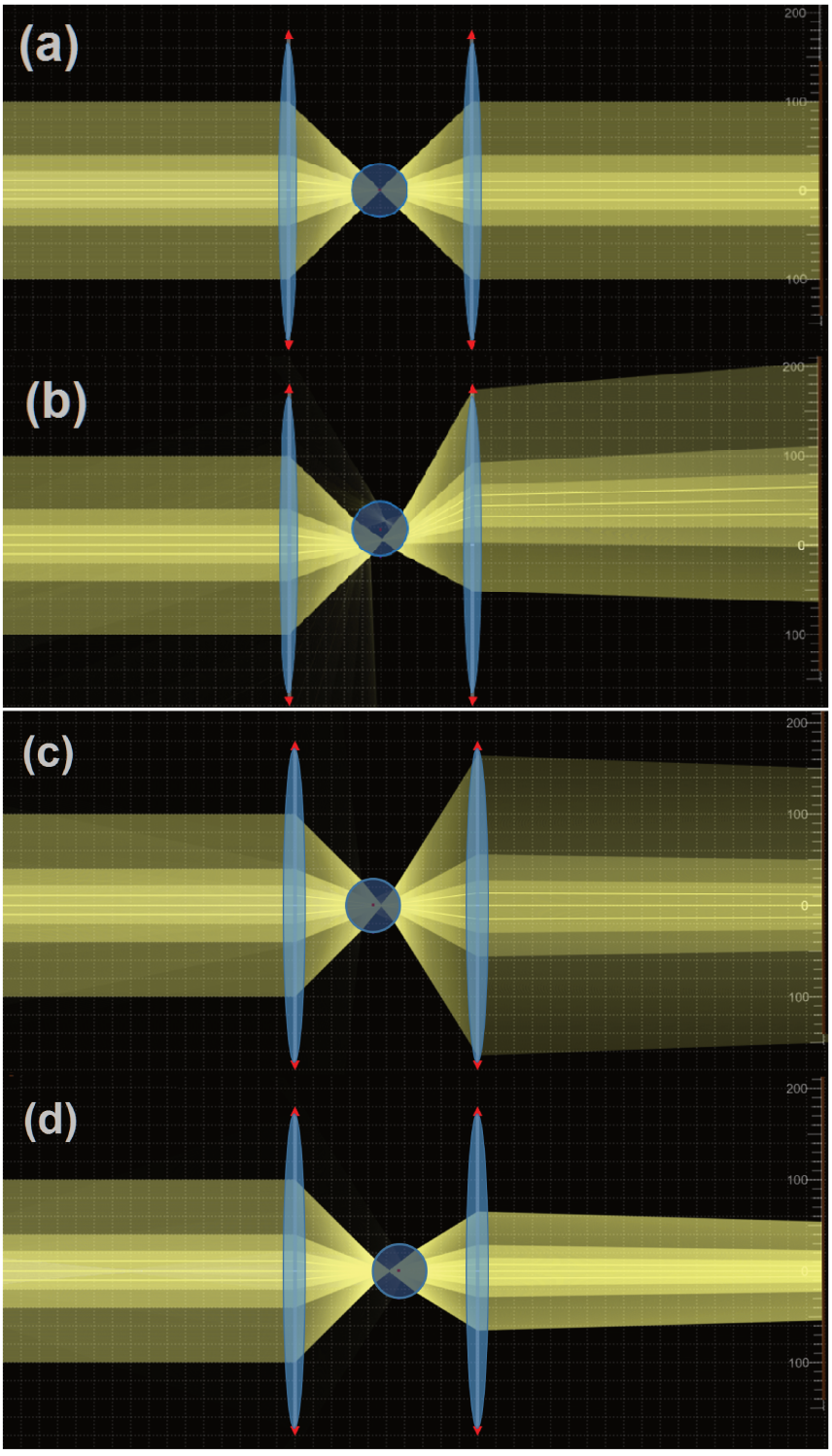


Figure 20: Effects on a focus laser beam of the displacements of a rigid transparent particle from the focal point: (a) rest position; (b) an orthogonal displacement from the optical axis break the cylindrical symmetry and the maximum intensity of the beam after the interaction with the particle is moved; (c) an axial displacement closer to the beam incomming direction refracted the beam in a larger area; (d) an axial displacement behind the focal point reduce the enlighted area.

3.3 THE MINI-TWEEZER

Optical tweezer is the name given to the class of scientific instruments that, using the optics principles developed in the previous Section, experimentally realize a controllable optical trap. During the collaboration with the Small Biosystem Lab in Barcelona, I had the opportunity to work with the version of the instrument called *mini-tweezer*. This name was given by its inventor, Steve Smith of the California University at Berkeley, when he developed the miniaturized evolution of previous instruments being used in the laboratory of Carlos Bustamante, one of the pioneers of this technique. This design uses two microscope objectives in a configuration that reproduces the optical setup before described, and uses two laser beams to compensate the *scattering force*.

The complete instrument is composed by:

- the tweezer housing, where the optical path is confined, including the laser diodes, fiber optics, the entire optical circuitry with the objectives, sensors and the CCD camera;
- the microfluidic chamber, where the experiment actually takes place; it is mounted inside the tweezer housing by a suspension support, with syringes attached by plastic microtubes where the solution can be manipulated by the experimenter;
- the laser controller, to control laser temperature and intensity;
- the electronic controller, directly connected with the tweezer; it controls the motors, the wigglers, and collects the information from the sensor by converting the analog signal into a digital information;
- the host program, a software installed on a computer that exchanges information in real-time with the electronic controller; it can also pilot the motors and the wigglers while collecting data;
- a monitor screen, where the images from the CCD camera are shown to directly inspect the chamber.

The mini-tweezer is suspended via a spring support, to reduce the vibration from the environment (see again Fig. 21). Moreover, during the experimental run it is covered by an acoustic insulator box that protects it also from external parasitic light, and contains temperature fluctuations. Finally, the mini-tweezer is connected with the laser power source and the electronic controller, and the experiment can start. More details of the experimental set up are given in Appendix A.

The forces that is possible to apply with the mini-tweezer lie in the range ~1 to ~100 pN. This is a rather wide range of forces that is particularly suited to follow in real-time nucleic acids and proteins through the folding/unfolding transition. The acquisition data-rate

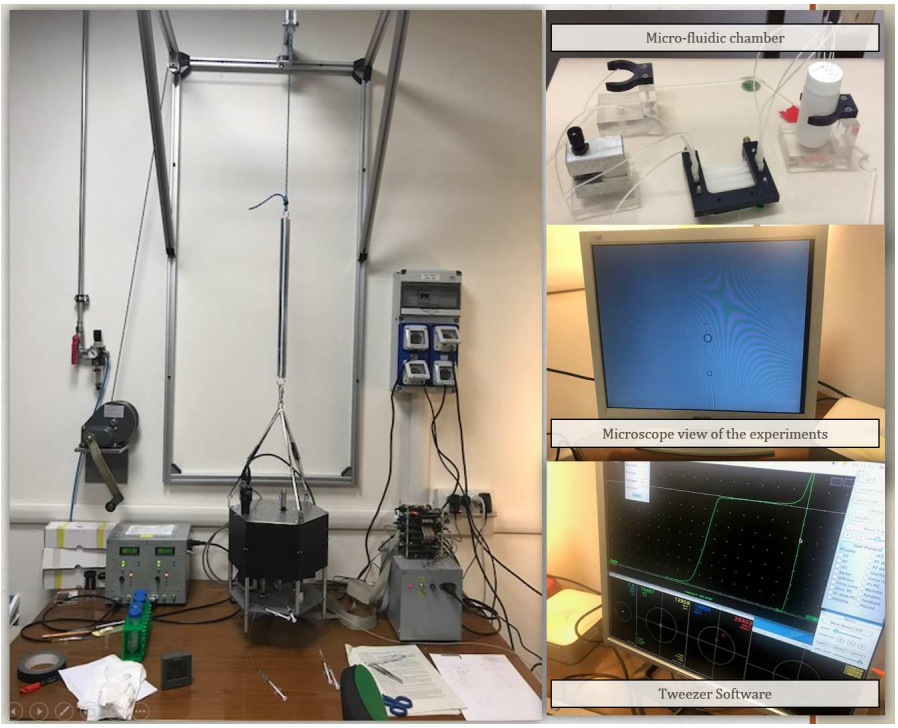


Figure 21: The Mini Tweezer from the Barcelona laboratory. On the left, the very compact instrument with all its parts occupy no more than a desk. On the right, from the top: the microfluidic chamber mounted on the support, connected with syringes, and the fluids trash; the image of the CCD camera where it is possible to observe the micropipette trapping one of the beads via air suction, while the other is captured by the optical trap (not visible); the custom software showing the force extension curve of λ-DNA.

for the host software is set at 1 kHz but could be increased (up to ~20 kHz) if the data are collected directly from the electronic controller.

The raw information that is obtained from this instrument is the position of the two laser spots on the focal plane, and the force applied from each laser on the trapped micro-bead. A critical issue concerns the perfect alignment of the two optical axis of the lasers, and the maintaining of a common focal point for the two objectives, for the entire duration of an experiment, allowing only for minor adjustments during each run. Careful calibration procedures exist for the force and position, to ensure the correct conversion of the variation of light collected on the sensors, to a corresponding variation of the force and displacement. These are carried out by using reference beads of known characteristics.

3.4 HAIRPIN FREE-ENERGY LANDSCAPE

Using the simple polymer models described in the first part of the Chapter, it is possible to reconstruct the force-distance curve of a complex polymer in terms of simpler quantities, such as: number of opened/closed base-pairs, actual hairpin length, number of base-pairs for each handle, and so on.

First of all, by using the energy contribution of each base-pair as obtained from the *nearest-neighbors* (NN) model, we can provide a theoretical estimate of the *free-energy landscape* (FEL) of the hairpin, and some other important properties. In general, the FEL is a multivariate function depending on the many internal degrees of freedom of the molecule. During the manipulation in the optical tweezer, its complex unfolding path is projected on the generalized coordinate that describes the molecular end-to-end distance. Given the measurement time orders of magnitude longer than the molecular fluctuation time, for sufficiently slow process the resulting FEL is an average of all possible microstates compatible with the total molecular extension corresponding to each value of the generalized coordinate.

It can be shown [19, 112] that DNA hairpins are correctly characterized by the number of opened base-pairs, a quantity which is related to the equilibrium molecular end-to-end distance (via the model used to describe the polymer extension). Coherently with the polymer models elaborated in the Section 2.1, the free-energy $\Delta G(f,n)$ of the hairpin with n opened base-pairs at a given force f, is given by the sum of:

- the free-energy at zero force ΔG^0_n, which can be estimated by the NN model;
- the reversible work done to orient the hairpin folded double-helix along the direction defined by the force, $\Delta G_{orient}(f,n)$ is obtained from the Eq.(2.5) for a single monomer orientation when $n = 0,\dots,N-1$ monomers, and no contribution when $n = N$;
- the reversible work needed to stretch the already opened hairpin base-pairs, $\Delta G_{ssDNA}(f,n)$, which may be estimated from the WLC model.

In particular, for the last term we can write:

$$\Delta G_{ssDNA}(f,n) = (1-\delta_{0n})(2-\delta_{Nn})\Delta G_{WLC}(f,n) \tag{3.11}$$

where $\Delta G_{WLC}(f,n)$ is the WLC contribution to ssDNA stretching, obtained from Eq.(2.10) for a chain with total length $L_0 = nb$ if $n < N$ or $L_0 = b(2N + N_{loop})$ for $n = N$. The Kronecker's deltas ensure that such a contribution is zero when the whole hairpin is in the folded state; instead, two equal contributions from the two ssDNA chains are added when the hairpin is partially opened, and the entire ssDNA sequence (including the loop) is accounted as one single chain when the hairpin is the fully open state. The total difference is therefore:

$$\begin{aligned}\Delta G(f,n) \quad &= \Delta G^0_{NU} + (1-\delta_{Nn})\textstyle\int_0^{f_N} x_{orient}(f)\,df + \\ &+(1-\delta_{0n})(2-\delta_{Nn})\textstyle\int_0^{f_U} x_{WLC}(f)\,df\end{aligned} \tag{3.12}$$

A few notes are in order: (1) the expression for $x_{WLC}(f)$ is purely formal, since we actually know its inverse $f_{WLC}(x)$ (see Eq.2.9), however the monotonicity of the latter assures the possibility of inverting

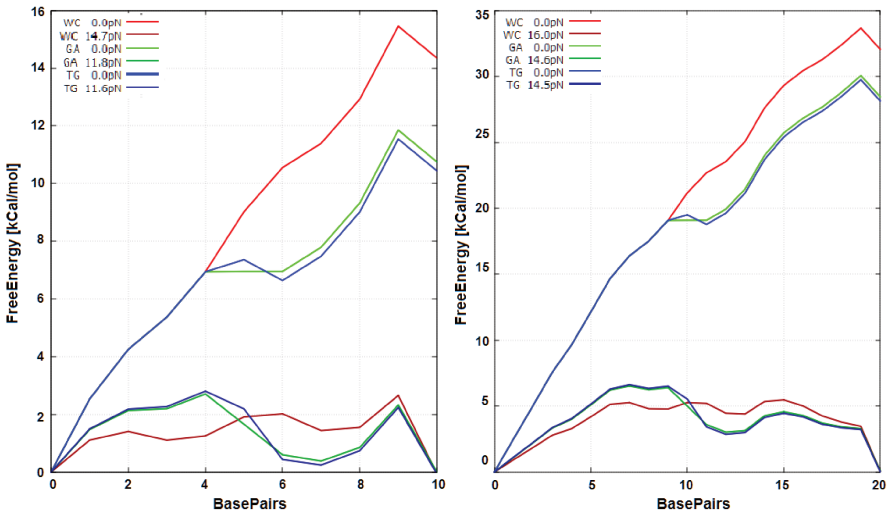


Figure 22: Theoretical prediction from the nearest-neighbor model of the free energy profile at constant force for the two groups of sequences: on the left the 10bp hairpin in native, GA, and GT configurations; on the right the same for the 20bp hairpin. For each configuration two sets of curves are represented: at zero force (lower branches), and at the respective coexistence force (upper branches, see text for details). Note that the small 10bp hairpin sequence in the left panel should correspond to the central 6-15 bp within the longer 20bp hairpin sequence in the right panel. .

it, albeit numerically; (2) when the force is fixed the contribution to the reversible work done by the trap and the DNA handles is equal for all hairpin states, so these contributions do not influence the equilibrium state; (3) moreover, application of the WLC in this context faces its limits for very small n, since one basic hypothesis for its validity is that the total chain length is longer than the monomer size b.

In hairpin sequences for which the Chargaff (or Watson-Crick) pairing rules are respected, the term ΔG^0_n is monotonically decreasing for $n \to N$ since breaking the H-bonds between the base-pairs costs energy; however, opening of the last base-pair is energetically favorable because it is accompanied by release of the torsional energy of the loop (not considered in the NN model). If we consider the overall energy profile as the sum of all contributions, it is possible to determine the coexistence force f_c (or *critical force*) at which the free-energy difference between the two states is zero. At this mechanical equilibrium condition, as seen in the Section dedicated to the transition-state theory, the two states are equiprobable, so they are observed with the same frequency.

In our experiments we studied six different DNA hairpins, divided in two sets with different stem length, respectively, of 10 and 20 basepairs. The particular base sequence was chosen so as to have an equal concentration of C-G and A-T basepairs to avoid the presence of local minima in the free-energy profile that would introduce intermediate state during the unfolding/folding path. For each set, the "native"

species has a correctly paired sequence (see Figure 23, top row, in the next Section); two mutants were created for each set by synthesis, including the GA or the GT mismatch near the stem centre (see sites indicated in red in the lower rows of Fig.23).

The theoretical free-energy profiles of the native and defective hairpins estimated by the nearest-neighbor (NN) model, are shown in Figure 22. By looking at the higher-lying sets of curves (corresponding to the calculation at $f=f_c$ for each hairpin), it can be readily noticed the energy jump induced by the mismatch defect (blue and green vs. red curves), with an effect also on the immediate neighbors; for the lower-lying set (model calculations in the $f = 0$ limit of zero applied force), displaying the theoretical energy barriers, the insertion of the mismatch is clearly visible at the crossing point between the blue/ green and the red curves.

3.5 EQUILIBRIUM EXPERIMENTS: HOPPING

The ssDNA sequence used to build the hairpins is received from the supplier as a single unit composed of 29 + 24 (or 44) + 29 bases (sequences shown in Figure 23), with the biotin already attached at the $5'$ end. The digoxigenin is subsequently attached at the $3'$ end, by a series of tailing and purification reactions (Roche DIG tailing kit, Qiagen QIAquick nucleotide removal kit). The two dsDNA handles are subsequently created on top of this structure, attaching the complementary strands to the terminal 29-base sequences on each side, by a standard annealing reaction; the entire dsDNA-hairpin-dsDNA construct is short enough to avoid splitting into two oligonucleotides, so that no further ligase reaction at the free dsDNA terms was necessary.

All experiments were performed with buffer Tris EDTA pH 7.5, 1 M NaCl at room temperature (298 K). We also added a small quantity of sodium azide (0.01%) to reduce the photodamage effects on the trapped molecule. In fact, the acid will function as singlet-oxygen scavenger [88] that helps in reducing the radicals naturally produced in the surroundings of the optical trap due to the interaction between the laser and the solvent.

In **hopping experiments**, the relative positions of the micropipette and of the optical trap (the aperture parameter λ in Eq.(2.36), see Fig. 18) are kept fixed at a given value: the hairpin is free to occupy either the *folded* or *unfolded* state, eventually jumping ("hopping") from one to another, while information about the force applied to the beads is recorded. For coherence with the notation of the previous Chapter, the folded state will be labeled N, and the unfolded state U.

In each experiment, we collected information on the applied force at 1 kHz for 15-40 s, then moved slightly the distance λ and repeated the measurement. Notably, once λ is fixed, the force in the unfolded state is lower than that in the folded state, because the microbead experiences a much larger displacement from the centre of the optical trap when the hairpin is closed (note that at similar forces, the exten-

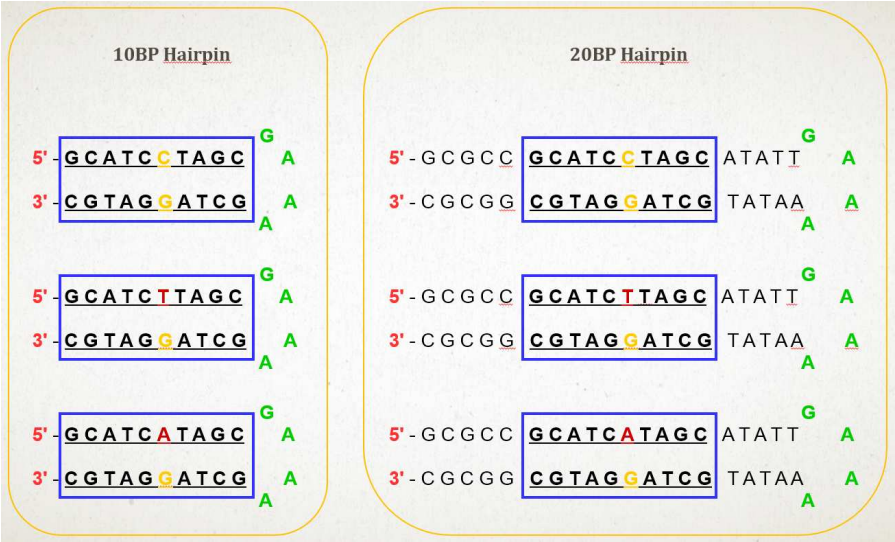


Figure 23: Sequences of the DNA hairpins used in the experiments. In all the sequences the unpaired loop is represent in green, while the base-pair where the mismatch may be introduced is highlighted in yellow (non-modified base) or red (modified base) On the left, the 10bp hairpin set; the total stem length is 20 bp, plus the 4 bases forming the loop. On the right: the 20bp hairpin set, including the 10bp part in the blue box with the same stem. For each sample, the two 29bp dsDNA handles are attached at the 5′ and 3′ hairpin ends.

sion of the unfolded ssDNA hairpin is much greater than the DNA helix diameter).

As consequence, if such a force difference between the two states is larger than the thermal noise, collecting information on the instantaneous force is equivalent to collecting information on the state (folded vs. unfolded) of the hairpin. Each variation of λ produces a consequent variation of the probability to observe the system in either one of the two states. Upon repeating the measurements for the collection of time-series taken at different λ's, we could measure the relative variation of the occupation probability of the two states. Figure 24 shows a two-dimensional histogram of the probability in the force-displacement Cartesian plane. In each plot it is possible to observe two close-by regions (colored ellipsoids). For each plot the distribution in positions for the two regions remains centered at a fixed average value (showing that except for thermal fluctuations the extension λ is actually constant), while the force has a different average values. This means that the two regions represent different observed states of the hairpin. In the panels from A to H, the extension λ (distance between the micropipette and the optical trap) is decreased and, as a direct consequence the equilibrium distribution probability moves from a state to the other. (It is to be noted that the λ in this plot is not directly representing the pipette-trap distance, but just the position of the trap; in fact, during the experiments the position of the microfluidic chamber could be slightly adjusted to maintain zero force in the $\hat{x}, \hat{z}$ plane and this movement is not taken into account in this plot.)

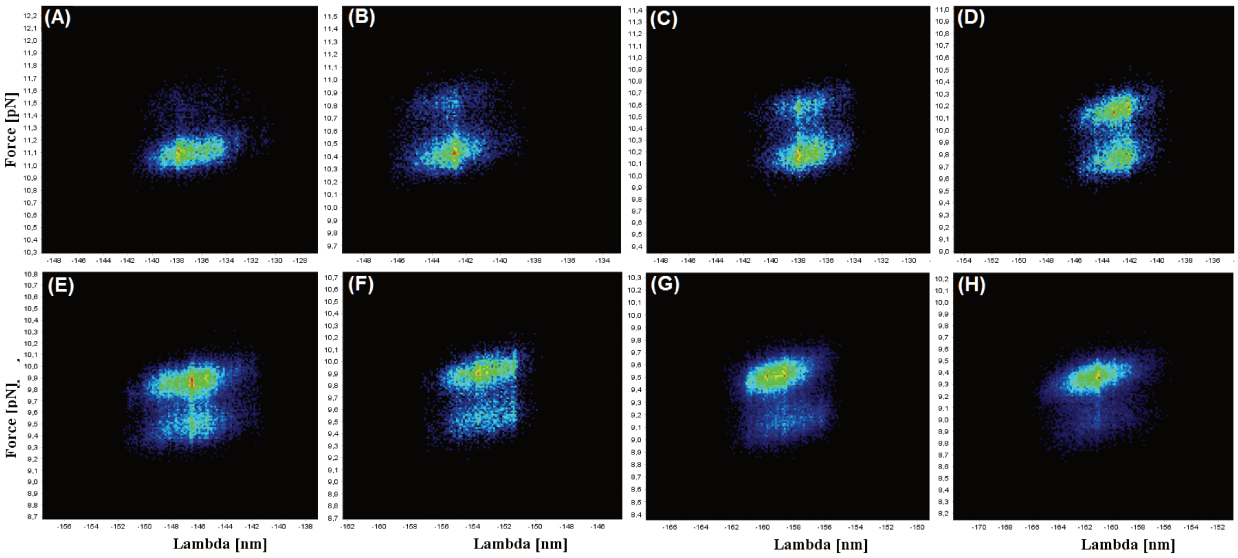


Figure 24: Two-dimensional histogram in the force-displacement plane during typical hopping experiments. The arbitrary color scale describes the population fraction compared to the maximum occupancy, from zero (dark blue) to 1 (red).

A key to interpret the optical tweezer experiments and to extract free-energies of bonding, is to combine the theoretical models from polymer physics with the concepts of the *reaction rate theory*, both described in the previous Chapter. Close to the coexistence value λ_c, at which the two states have equal occupation probability, the kinetic rate of unfolding (k^+) and refolding (k^-) fall in a timescale that allows to observe several hopping events Therefore, for each time-series taken at fixed λ, we could reconstruct:

1. the probability distribution of finding the hairpin in the folded/unfolded state (see histogram in Figure 25, top panel);
2. the average force in the folded/unfolded state;
3. the sequence of transitions along the time evolution (the green line in the time-series in the bottom panel of Fig. 25);
4. the transition probability between the two states (directly related to the reaction kinetic coefficients k^+, k^-).

Typical time-series and probability histogram data for the 10bp hairpins, with either native sequence or GA and GT mutated, are presented as multi-panel Figures 26, 27 and 28.

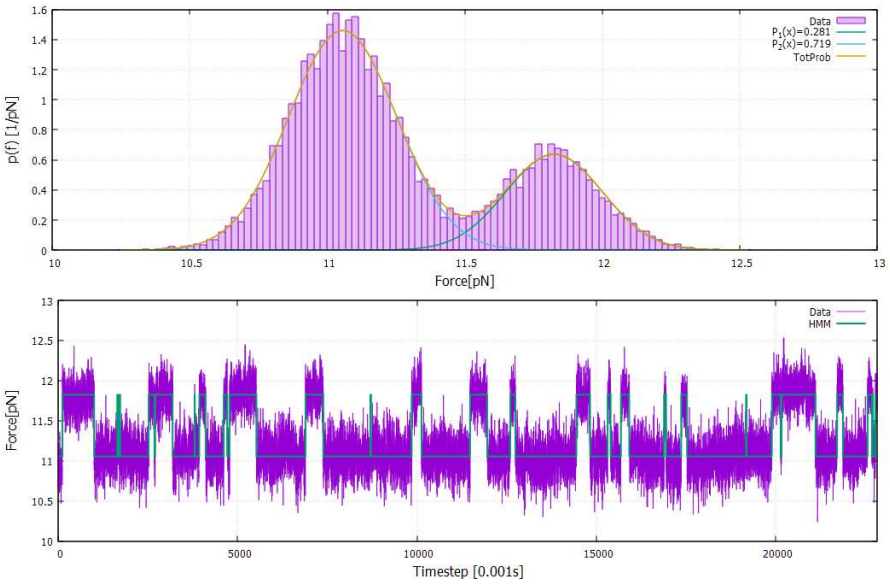


Figure 25: Example of a time-series of force measurements (bottom) from a hopping experiment of the 20bp GA hairpin; the time-series represents the recorded force as a function of the instrument time step ($dt = 0.001s$). The plot above is the histogram of the collected values of instantaneous force, together with the best-fit from the Hidden-Markov model (green curve); the two Gaussian peaks correspond to the force distributions in the folded and unfolded state.

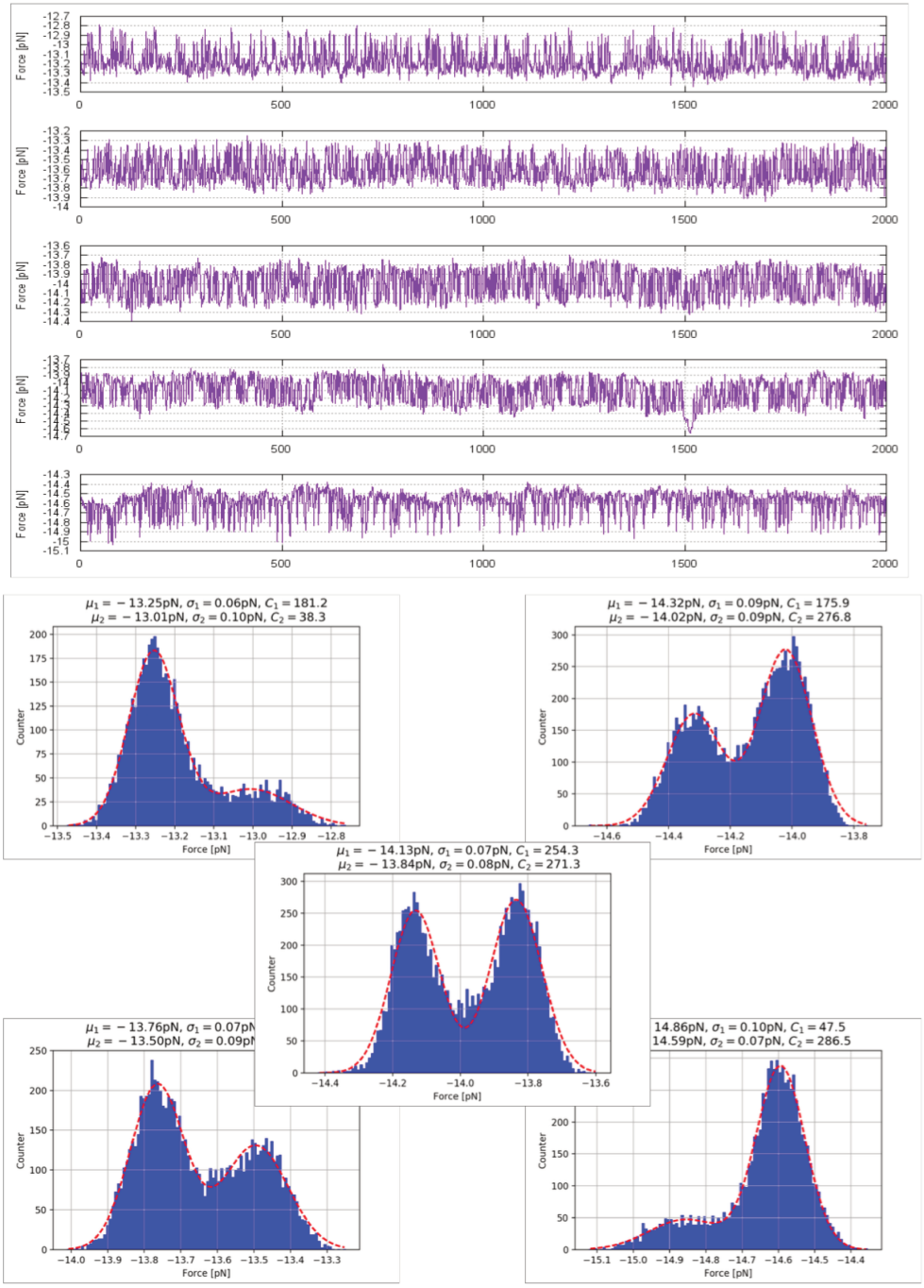
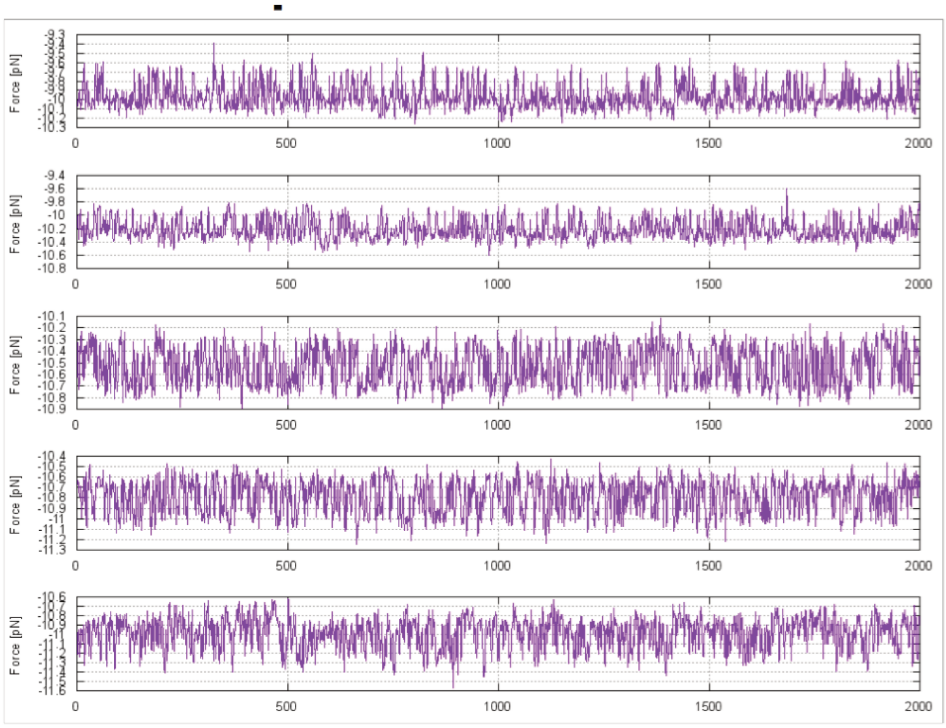


Figure 26: Experimental plots from hopping experiments on the 10bp hairpin with native (Watson-Crick) sequence, around the critical force $f_C = 14.$ pN. Above: force-time traces for five positions at increasing optical trap opening. Below: the corresponding force distribution histograms, the two maxima indicating the equilibrium forces corresponding to the unfolded and refolded hairpin conformation.



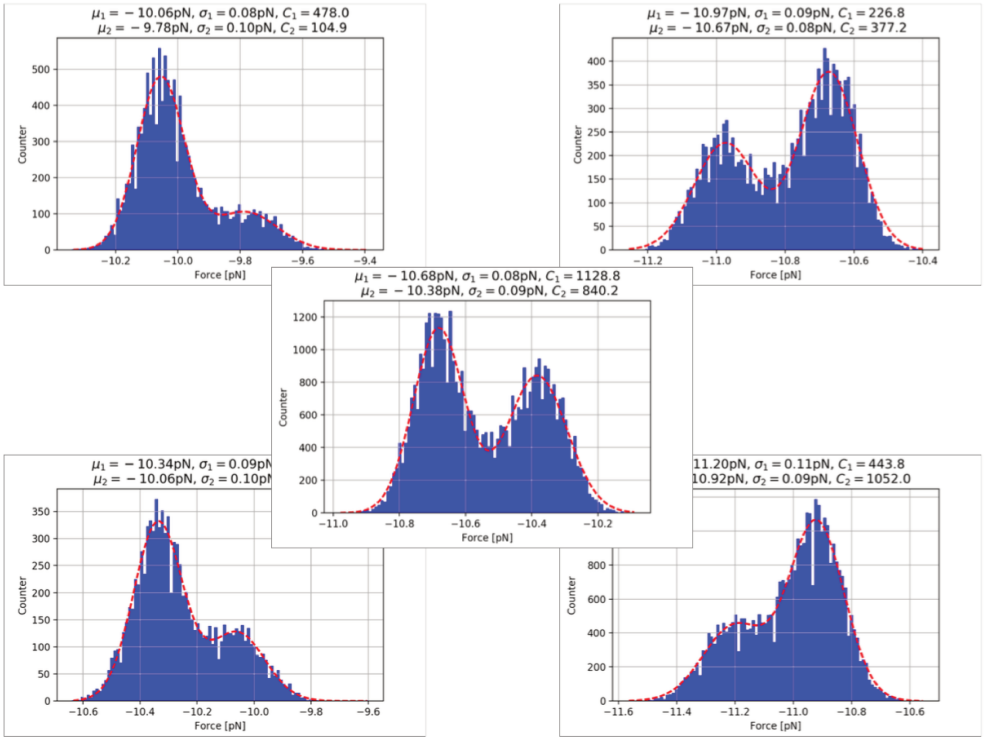


Figure 27: Same as Fig.26 for the 10bp hairpin with a GA mismatch defect, around $f_C = 8.6$ pN.

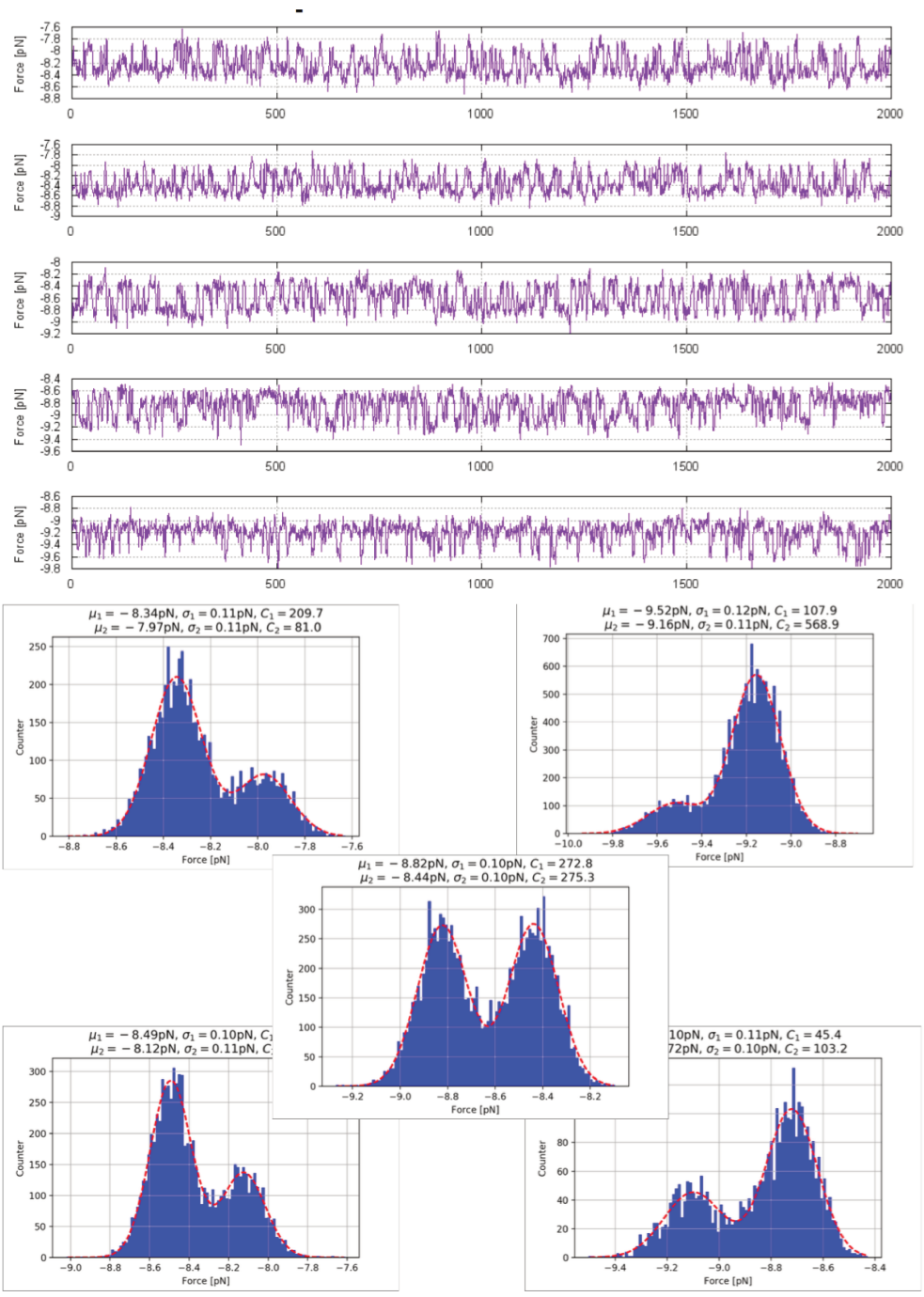


Figure 28: Same as Fig.26 for the 10bp hairpin with a GT mismatch defect, around $f_C = 10.5$ pN.

3.5.1 *Data analysis with hidden Markov model*

The data from the experimental time-series were analyzed with the Hidden Markov Model (more details of which are given in the Appendix B). This algorithm determines the likelihood of a given probability distribution for a set of hidden states with respect to a finite Markov series of observations: in our case, the fluctuating force at fixed displacement. Using an optimization process it is possible increase the likelihood of the probability distribution in such a way to fit the hidden states, and the probability of transition between such states. Each element in the series of measurements has a certain probability to belong to a particular state, depending on the values of the measurement itself, and the probability distribution associated with the state.

We start from the tentative hypothesis that the system can only be in two states (N and U), and that each state probability distribution is Gaussian. This assumption is justified *a posteriori* by looking at the distribution histograms of the type shown in Figs.25-28. Indeed, no "hidden" states beyond the two folded/unfolded configurations emerged from that analysis. Therefore, from the optimized parameters after the HMM algorithm, important information on the folding/unfolding reaction can be obtained :

- the average force $\langle f_U \rangle$ and $\langle f_N \rangle$, and its standard deviation, in the two states (that is, the mean value and FWHM for each state's Gaussian distribution);
- the kinetic coefficients k^+, k^- of the folding/unfolding process, those are given by the off-diagonal values of the transition matrix T divided by the timestep of measurements collecting rate;
- the probability w_U, w_N to observe either state during the measurement process, at a given opening λ.

It must be noted that for any fixed value of λ, the two kinetic coefficients correspond to generally different values of force, namely: (i) the value f_N of the force acting in the folded state for k^+, and (ii) the force f_U acting in the unfolded state for k^-. Therefore, we must separately describe the variation of $k^+(f_N)$ and $k^-(f_U)$ as a function of the respective force values, with λ acting as a parameter. These data, collected in the semi-log plot of Figure 29, mark two important points:

- the exponential dependence of the kinetic rates on the force indicates that the Bell-Evans model should be readily applicable in this case;
- the coexistence force f_c is readily identified by the value at which $k^+ = k^-$ (see Table 1).

Furthermore, it is also proved that the two-state hypothesis is verified for the hairpin folding/unfolding transition and, with the parameters optimized from the HMM, the equilibrium probability dis-

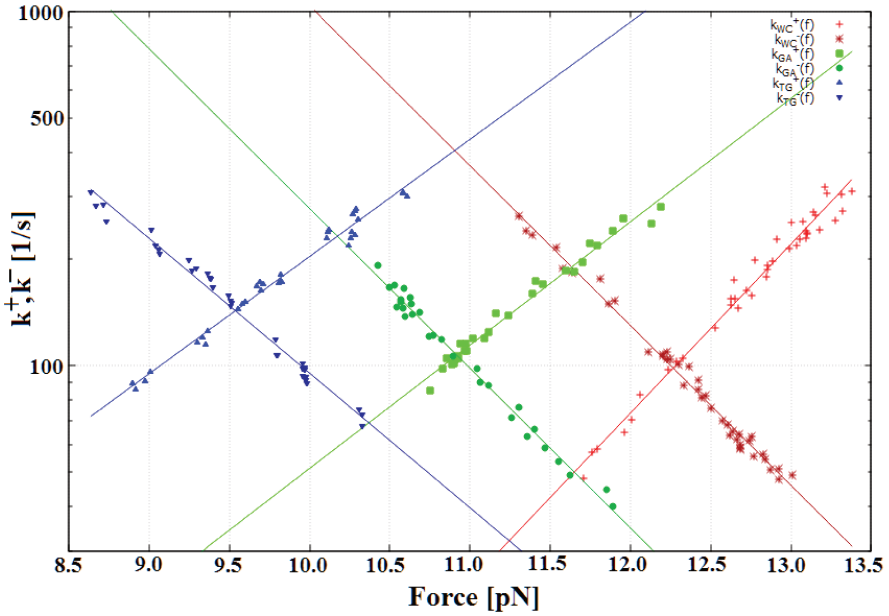


Figure 29: Plot of the unfolding kinetic rate k^+ (dark symbols) and folding kinetic rate k^- (bright symbols) as a function of the force, in hopping experiments for the 10bp hairpins: native (red), GA-mutant (green), and GT-mutant (blue). The points are the data extracted from the HMM analysis, and the lines are the fit with an exponential curve $k(f) = \exp(af + b)$, in agreement with the Bell-Evans model. The crossing point between each pair of folding/unfolding fits determines the coexistence force f_c, where the folded and unfolded state of the hairpin have the same occupation probability.

tribution between the two states can actually be written as a double Gaussian distribution:

$$P_\lambda(f) = \frac{w_{\lambda,U}}{\sqrt{2\pi\sigma^2_{\lambda,U}}} e^{-\left(\frac{f-\langle f_{\lambda,U}\rangle}{4\sigma_{\lambda,U}}\right)^2} + \frac{w_{\lambda,N}}{\sqrt{2\pi\sigma^2_{\lambda,N}}} e^{-\left(\frac{f-\langle f_{\lambda,N}\rangle}{4\sigma_{\lambda,N}}\right)^2} \tag{3.13}$$

where the two maxima represent the average force in either the folded or unfolded state, and the standard deviations the corresponding thermal fluctuation of the force.

These same values of average force $\langle f_U\rangle$ and $\langle f_N\rangle$, and the relative probability weights w_U, w_N, can be readily used to recover the free energy barrier of the folding/unfolding transition ΔG_{NU} from hopping experiments by using the Eq.(2.12) established in the previous

Table 1: Values of the coexistence force f_c in pN for hopping and pulling experiments, as obtained from the crossing point of the exponential fit (see Fig. 29, Fig 39 left column).

Hairpin	10bp hopping	20bp hopping	20bp pulling
native	12.3	15.9	15.8
GA	10.9	(15.9)	14.7
GT	9.5	14.1	14.5

Chapter, by noting that the ratio n_U/n_N practically coincides with the ratio w_U/w_N, at fixed λ.

Different contributions to the polymer model for the hairpin in the folded and unfolded state are used to describe the free energy of the system, and hence recover its zero-force limit [2]:

(i) the free-energy contribution to orient the DNA helix in the folded state, Eq.(2.5) with hairpin diameter $d = 2.0nm$;

(ii) the contribution to stretch the ssDNA chain in the unfolded state, Eq.(2.10) with the conventional persistence length $\xi = 1.35$ nm, and a fictitious nucleotide unit length of 0.585 nm (both values coming from a best-fit to the WLC curve for ssDNA in short hairpins [25]).

From the Eq.(3.12) for the total free energy, taken at the two extremes $n = 0$ (state N, folded hairpin, no open base-pairs) and $n = N$ (state U, unfolded hairpin, all base-pairs opened), the zero-force limit of the free-energy of the folding/unfolding transition is then [2]:

$$\begin{aligned} \Delta G^0_{NU} = \; & -k_B T \ln\left(\frac{w_U}{w_N}\right) + \int_0^{\langle f_U\rangle} x_{ssDNA}(f)\,df \\ & - \int_0^{\langle f_N\rangle} x_{orient}(f)\,df \end{aligned} \tag{3.14}$$

The results of this analysis carried out on several molecules for the two sets of hairpins (10 and 20 bps) are shown in Fig.30 and Table 2, together with the corresponding theoretical predictions of the NN model (that is, the integrals of the zero-force curves in Fig. 22). It may be noticed that the NN model predictions are systematically larger than the experimental values, in some cases by up to 20%, a possible explanation of this discrepancy has been found comparing the the ideal model with the molecular dynamic simulation of the same 10 bp hairpin sequence.

Table 2: Free-energy differences (kcal/mol) between the folded/unfolded state for the 10bp and 20bp DNA hairpins from hopping experiments, and corresponding theoretical prediction from the nearest-neighbor (NN) model. Data for the native configuration, and including a GA or GT mismatch defect. The last column "NN-corrected" refers to the model value minus the first base-pair, a procedure whose reason is explained in Chapter 5.

Hairpin	ΔG	Error	NN	NN corrected
10bp native	12.5	0.9	14.35	11.82
10bp GA	8.8	0.6	10.75	8.22
10bp GT	9.1	1.0	10.44	7.91
20bp native	29.8	0.1	32.05	29.52
20bp GA	27.6	0.2	28.45	25.92
20bp GT	25.5	0.4	28.14	25.61

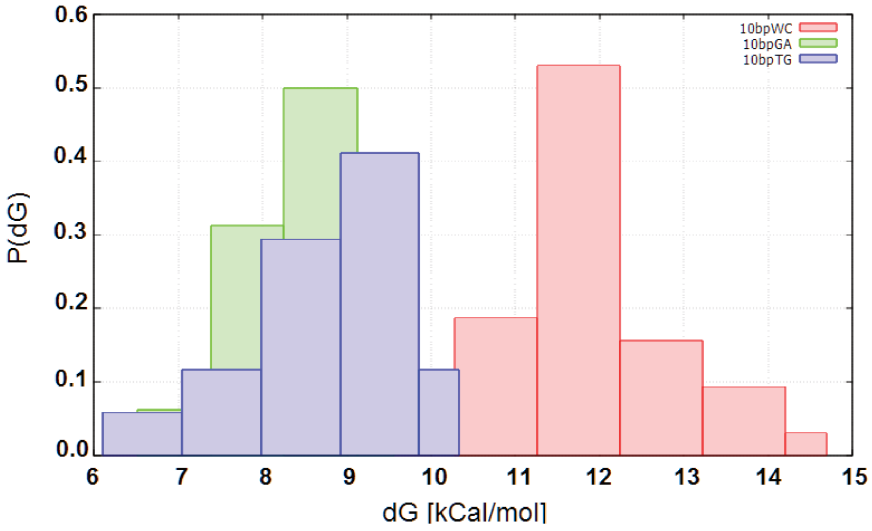


Figure 30: Histograms of the free-energy ΔG_0 obtained from the hopping measurements on the 10bp hairpins: native (red), GA-mutant (green) and GT-mutant (blue).

It must be noted that the transition rate for the longer 20bp hairpins is relatively small in the experimental time window, i.e. very few folding/unfolding transitions are observed, compared to the large number of "hops" recorded for the 10bp hairpins. This implies a rather large error in the kinetic rate estimates for the longer hairpins, for which the force-pulling method turns out to be more suitable (see next section; the apparently small error quoted in the Table for the 20bp is simply due to the very small number of of molecules for which we could obtain significative statistics).

3.6 NON-EQUILIBRIUM EXPERIMENTS: PULLING

In constant-velocity pulling experiments, the position of the optical trap λ is moved at constant speed, sweeping back and forward between two fixed force values. As a consequence of the extension variation, the force applied to the DNA hairpin also increases/decreases during the pulling cycle. As predicted by the Bell-Evans model, it is possible to move the equilibrium occupancy distribution between the unfolded and folded states:

- at low forces hairpins are in the folded state, with the stem forming a double helix;
- at large forces they unfold in a stretched conformation, where the entire hairpin (stem+loop) is found as ssDNA [104].

The force-displacement plots are collected at two-dimensional histograms of points, which are colored according to the frequency of occupancy of each hairpin conformation. Typically, two branches of force are observed (Figure 31): the upper branch shows the elastic response of the whole molecular construct when the hairpin is folded, whereas the lower branch shows the same response for the unfolded

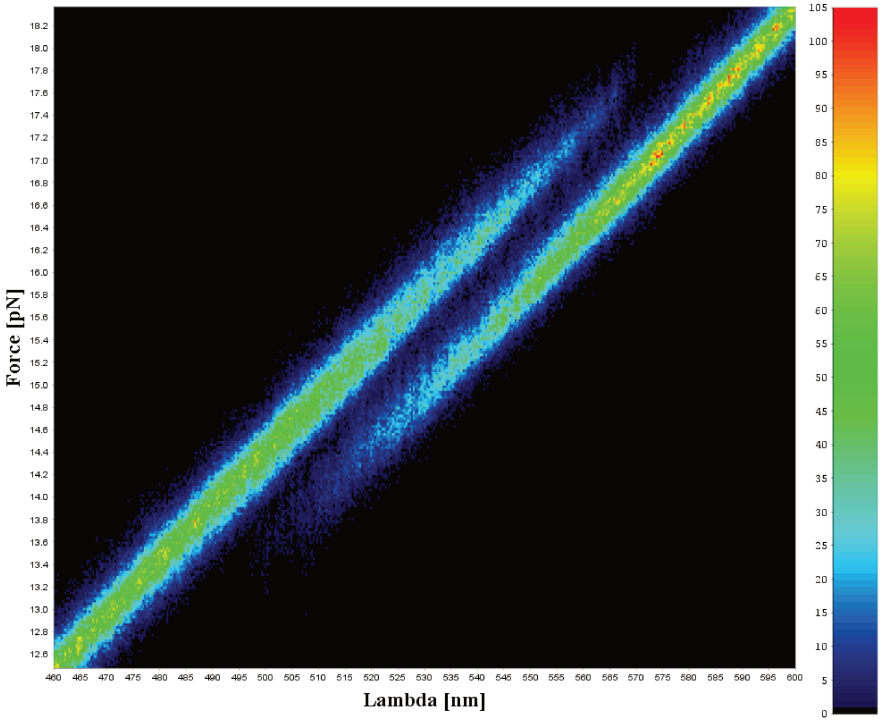


Figure 31: Histogram in the Lambda-Force plane during pulling experiments, similar to the hopping 24 the chromatic describe the position compare with the maximum occupancy in the plot, from zero (black) to one (red). In the plot are accumulated points from several cycles of the same *native* 20bp molecule at pulling speed 100nm/s. It is possible to observe the two branches at low forces the hairpin is in its natural state (folded) and increasing the force the equilibrium is moved until the hairpin is always in the open state (correspond to the lower force branch). Notice that in this histogram lambda is the represent the pipet-trap distance because even if the motor position between cycles is displace a process of *trajectories alignment* C is executed before evaluate the histogram.

hairpin. The folded and unfolded state appear as roughly diagonal bands clustering conformations belonging to each state. Transitions between the two hairpin states are viewed in the force-displacement plot as a sudden jump in the force. The critical force at which such transitions take place change upon repeated cycling of the same experiment, due to thermal fluctuations, displayed in the plots by the partial superposition of the two branches.

A series of *force-pulling* cycles of pull/release are performed at constant velocity. In our experiments we collected data at pulling speeds v_{pull} =50, 100, 150, 200, 300 nm/s (mostly between $50 - 100$). By varying the pulling speed, it is possible to increase/decrease the hysteresis between the folding/unfolding trajectories, thereby reducing the number of jumps in a single trajectory.

3.6.1 *Potential energy landscape from probability density*

One problem arises from the fact that information on the position and the force are measured on the entire molecule+beads assembly, but

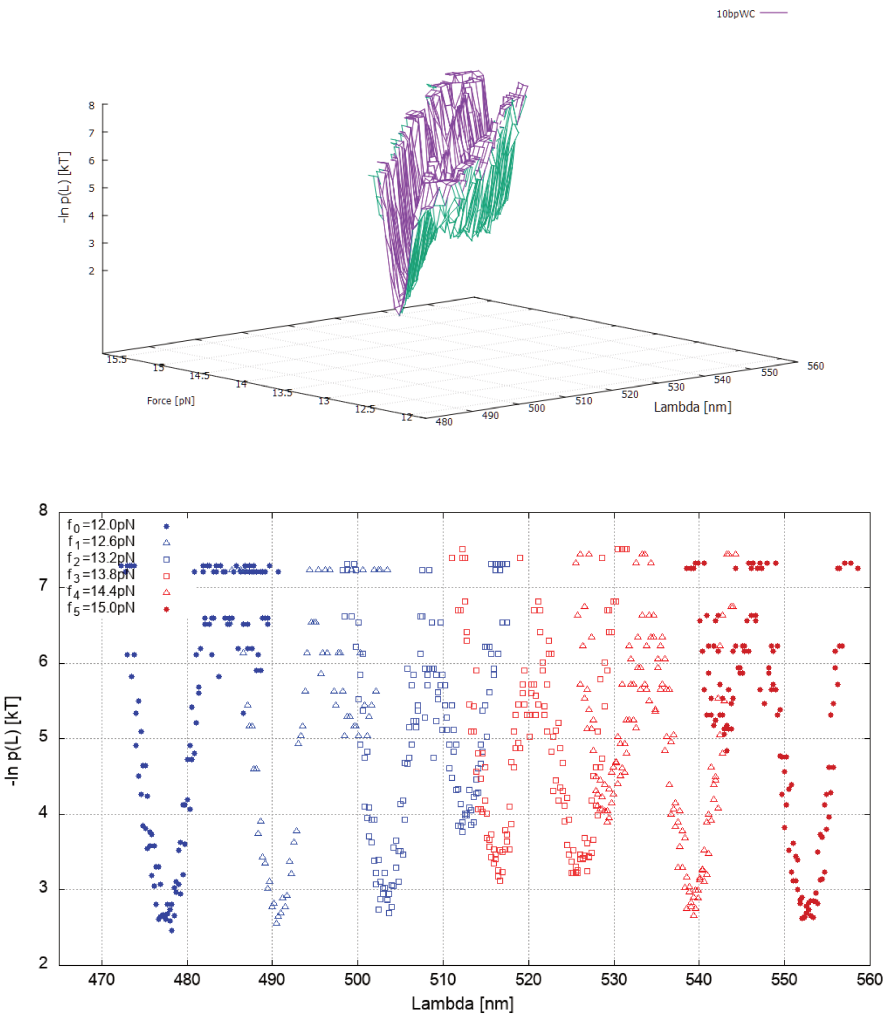


Figure 32: 3D Plot of the logarithm of the probability of finding the 10bp *native* hairpin at a given extension λ at different forces (top). In the lower plot 2D representation of the same $-\ln p(x)$ at different forces, from blue to red the different shade denote an increasing external force applied to the system.

we need to extract force and displacement relative to the hairpin only. For example, the experimentally obtained, raw probability density $p_f(\lambda)$ as a function of the total displacement λ for the 10bp hairpin is shown in Figure 32, where the quantity $-\ln p_f$ is plotted for a series of values of force around the coexistence; each set of points (differently colored) identifies a kind of double-well potential, that is however translating on the λ axis, upon increasing the force.

For the 10bp hairpins at constant velocity 50 nm/s the kinetics of the folding/unfolding reaction is extremely fast compared to the variation of the trap position (this is confirmed also by the analysis of the hopping experiments). Therefore, the hairpin is moving through quasi-equilibrium states, and the slow trap displacement does not distort the observed distribution of points in the force-extension plot. This is confirmed also by the almost identical distributions obtained by collapsing the data for the unfolding and refolding trajectories. For such quasi-equilibrium condition, the density $p_f$can be inverted, and

expressed as a function of the hairpin extension only (see Appendix D):

$$x_{ssDNA}(f) = \lambda - x_{H+B}(f) = \lambda - \left(\frac{1}{k_{fold}^{eff}} - \frac{1}{k^{d}(f)} \right) f \tag{3.15}$$

Once the probability distribution is recalculated as a function of $x = x_{ssDNA}$, applying the standard Boltzmann inversion procedure to the density allows to extract the profile of the effective potential surface $V'(x; f)$:

$$p_f(x) \propto \exp\{-\beta[V'(x; f)]\} \tag{3.16}$$

Interestingly, all the data for the effective potential are now aligned and superposed, In this way, the effect of the force on the potential energy landscape is nicely shown, and reproduces well the downward "rotation" of the potential profile by the external force, qualitatively already described in Figure14 of Chapter 2.

By repeating this analysis for the 10bp hairpin with TA and TG mismatch (Figure 34), it can be noticed the drastic lowering of the potential barrier, clearly signifying a less stable system compared to the native sequence.

3.6.2 *Free energies from the analysis of the first rupture force*

For the 20bp hairpins the kinetic rate is smaller and the transition time is larger, so that the quasi-equilibrium condition is no longer verified. Anyway it is possible to extract information from the trajectories by using other considerations, in particular the ones exposed in Section 2.2.4. By analysing the folding/unfolding trajectories, it is possible to detect the force value at which the hairpin unfolds, or refolds for the first time along the trajectory; this is called the *first-rupture force*, and appears in the force-displacement plots of the type shown in Figure 35 and Figures 36,37,38 (on the top row panels) as a nearly vertical step going from the upper to the lower line, or vice-versa (in all the plots, the hairpin folding trajectory is shown in blue, and the unfolding in red).

Repeating this procedure for many folding/unfolding trajectories at fixed pulling speed v_{pull}, gives the distribution of differential survival probability, $dP_N(f)/df, dP_U(f)/df$, by counting the histogram of first-rupture force (Fig.36,37,38 central row panels); and the corresponding survival probability, $P_N(f), P_U(f)$, as the fraction of trajectories that keep the initial state up to the force f. Using the equation obtained in the previous chapter

$$\begin{cases} \frac{dP_N(f)}{df} &= -\frac{k^{+}(f)}{R} P_N(f) \\ \frac{dP_U(f)}{df} &= -\frac{k^{-}(f)}{R} P_U(f) \end{cases}$$

This procedure is more effective in the longer hairpin where the jumps are well defined and in the order of 1-2 for each trajectories. In the 10bp hairpins it is possible to observe jumps even at lower force, with the consequence that rupture force distributions, for the

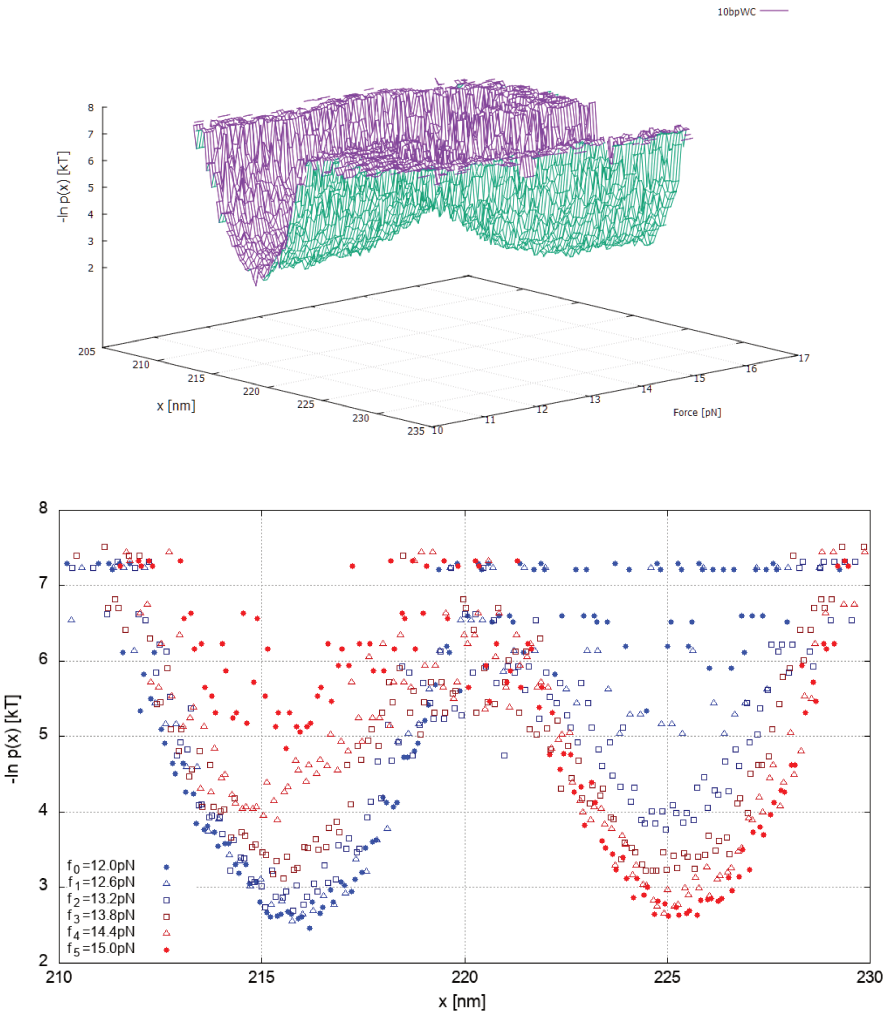


Figure 33: Plot of the logarithm of the probability of finding the 10bp native hairpin at a given extension x at different forces. In the lower plot, from blue to red the different color shading from blue to red corresponds to increasing external force.

unfolding and refolding trajectories, overlap only in the tails where the uncertainty due to the statistical sampling is greater. In theory this problem could be avoid increasing the pulling speed, but the precision in the position detection is reduced if the velocity is increase above a certain value because the protocol in the host software that manages the alignment of the two lasers is no longer able to maintain a continuous alignment, so for reason of time we prefer to continue this analysis only with the 20bp hairpins.

The parameter R is the constant-force pulling rate, determined by the pulling speed v_{pull}. In fact, close to the rupture force the force-extension relation may be approximated by a linear dependence. In this range the dominant contribution to the total stiffness of the system, with an effective spring constant K_{eff}, is given by the displacement of the bead within the optical trap, so that:

$$R = \frac{df(t)}{dt} = \frac{\partial f(x)}{\partial x}\frac{dx(t)}{dt} = K_{eff} v_{pull} \approx const \tag{3.17}$$

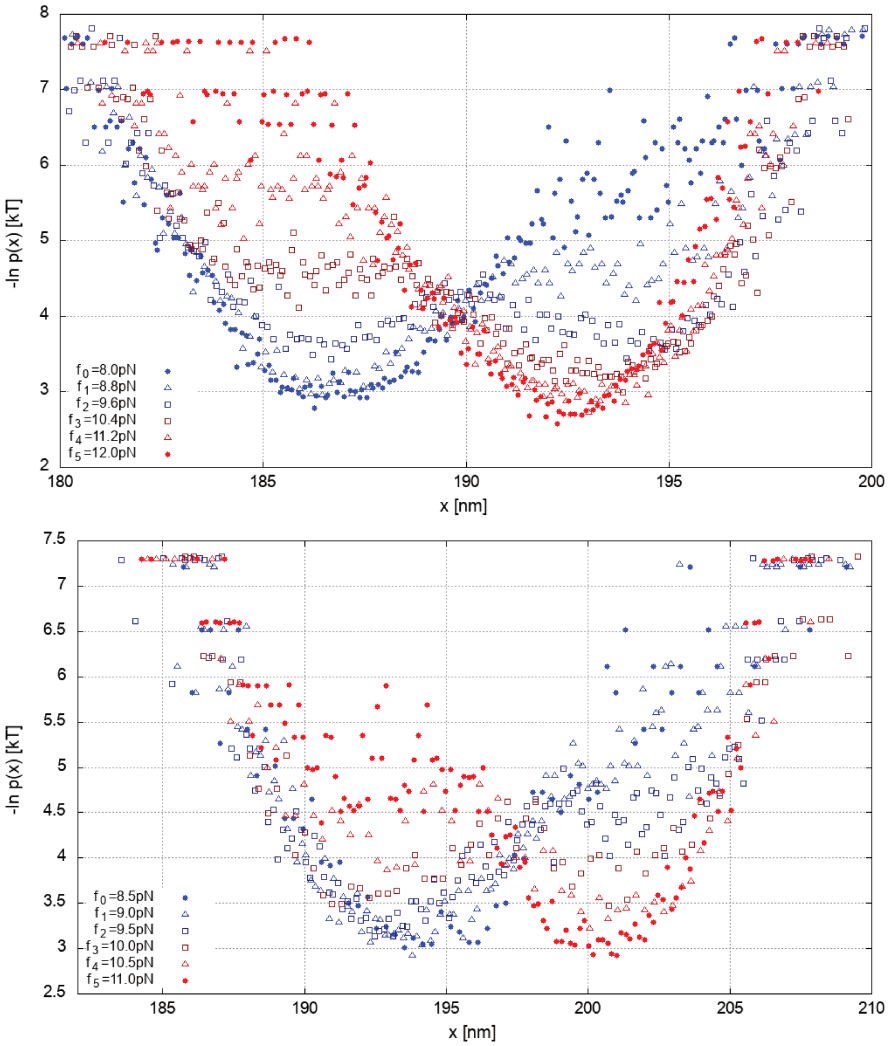


Figure 34: Same as Fig.33 lower panel, for the10bp hairpin including a GA (above) or a GT (below) mismatch.

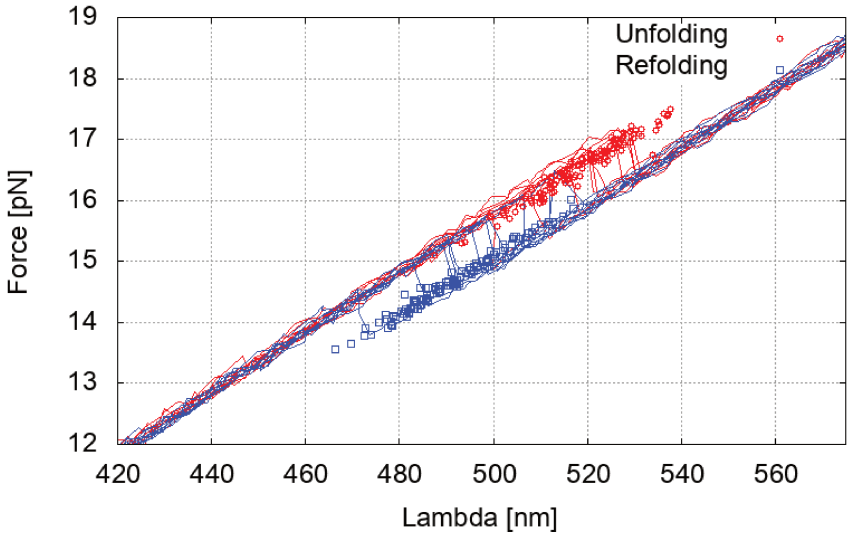


Figure 35: Trajectories of a force-pulling experiments for the native 20bp hairpin, at pulling speeds ν_{pull} =100. In the unfolding trajectories (red) are shown the firsts positions in which is observe a transition to the lower branch (first rupture force), likewise in the refolding trajectories (blue) are shown the firsts positions in which the system jumps into the upper branch (first refolding force, but we will generally refer to them also as ropture force). The rupture force corresponds to the nearly vertical jumps between each pair of trajectories.

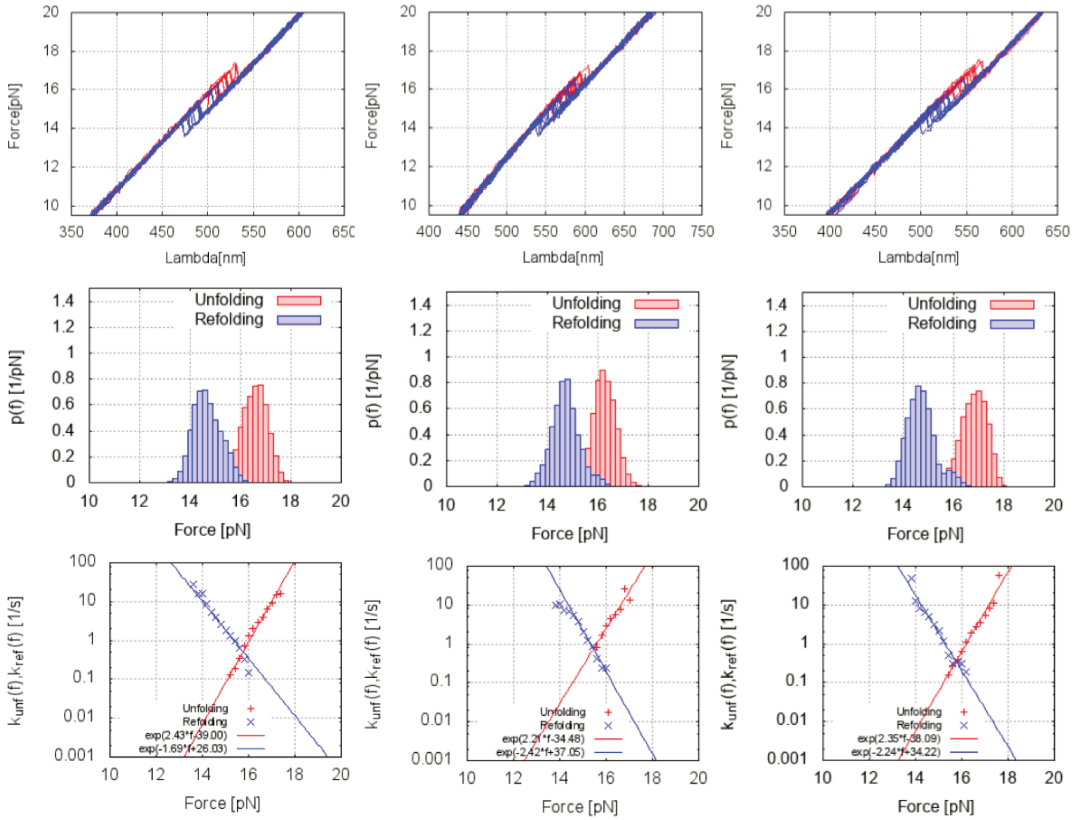


Figure 36: Experimental plots from force pulling experiments on the 20bp hairpin with native (Watson-Crick), at pulling speed 100 nm/s. Upper row: force displacement plot (blue trajectory for unfolfing, red trajectory for refolding). Middle row: the corresponding force distribution histograms. Bottom row: force-dependent reaction rates $k_U(f)$ (blue) and k_R (red).

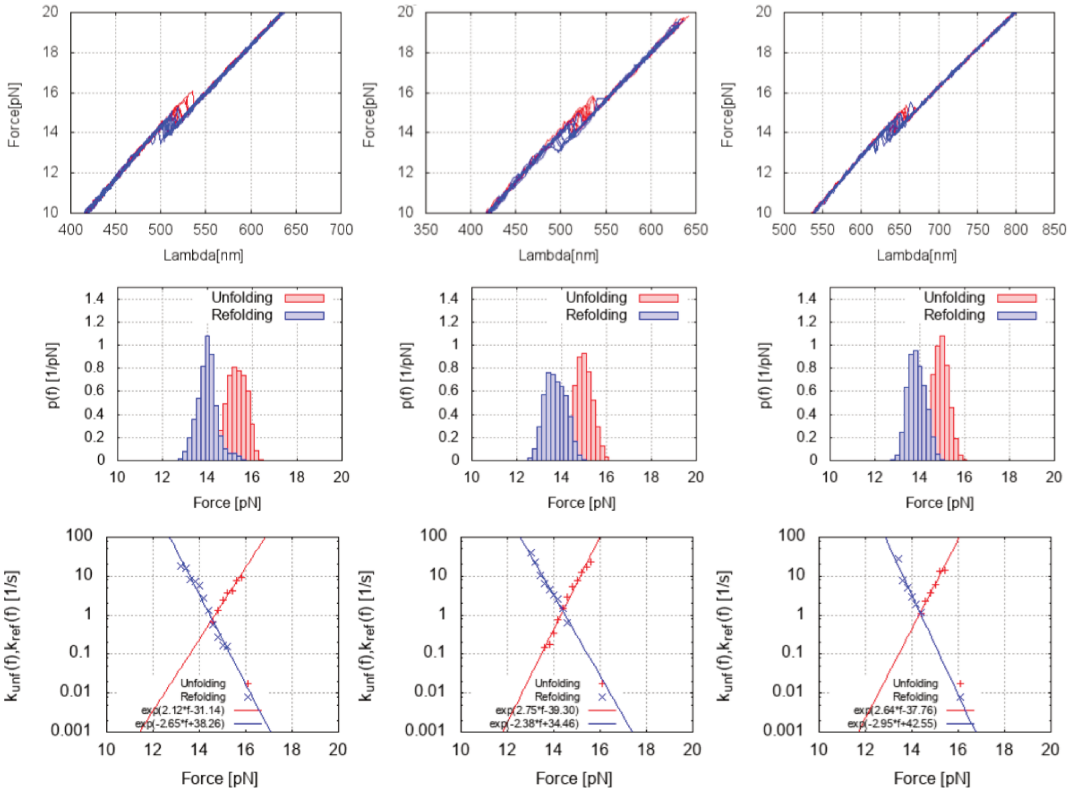


Figure 37: Experimental plots from force pulling experiments on the 20bp hairpin with GA mismatch sequence(top), at pulling speed 100 nm/s. Upper row: force displacement plot (blue trajectory for unfolfing, red trajectory for refolding). Middle row: the corresponding force distribution histograms. Bottom row: force-dependent reaction rates $k_U(f)$ (blue) and k_R (red).

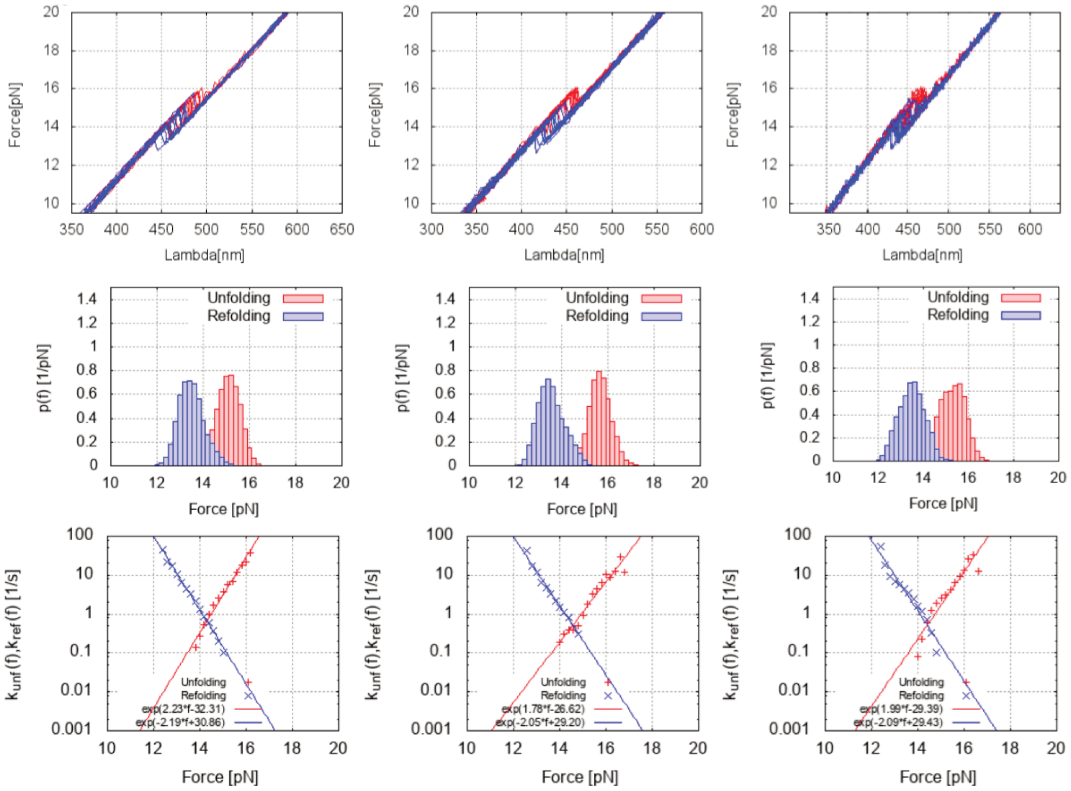


Figure 38: Experimental plots from force pulling experiments on the 20bp hairpin with GT, at pulling speed 100 nm/s. Upper row: force displacement plot (blue trajectory for unfolfing, red trajectory for refolding). Middle row: the corresponding force distribution histograms. Bottom row: force-dependent reaction rates $k_U(f)$ (blue) and k_R (red).

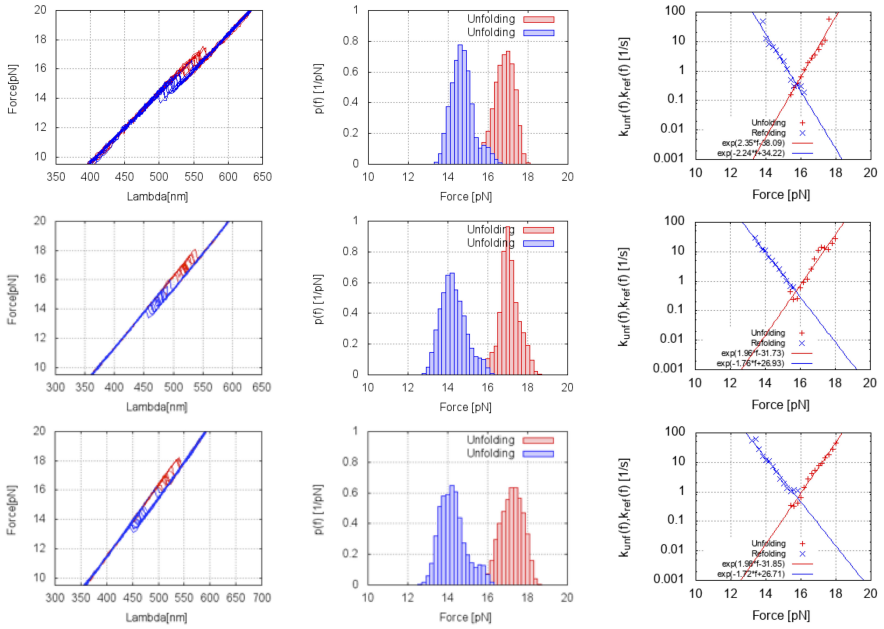


Figure 39: Summary of force-pulling experiments for the native 20bp hairpin, at different pulling speeds. From top to bottom, v_{pull} =100, 200, 300 nm/s. Left column: unfolding(red) and refolding (blue) trajectories; each experiment corresponds to about 100 cycles; the rupture force corresponds to the nearly vertical jumps between each pair of trajectories. Central column: histograms of the first-rupture force for unfolding (red) and folding (blue) trajectories; for clarity, the data are convoluted with a Gaussian smearing function with σ =0.2. Right column: unfolding (red) and folding (blue) kinetic rates, k_U, k_N; crosses are the experimental data, straight lines are the exponential fit according to the Bell-Evans model.

The transition-state theory developed in section 2.2, and in particular Eq. (2.21) and (2.22), describes the variation of the kinetic rate as a function of the force-dependent potential barrier, $E_B(f)$, that the system must overcome. By assuming that under the constant force the system spans a unidimensional reaction coordinate x, in the Bell-Evans model approximation the height of the potential barrier at the transition state point (x_B) depends linearly on the applied force (see Eq. (2.20) in Chapter 2).

The kinetic rates k_U, k_N extracted according to the Bell-Evans model for the native hairpin are shown in Fig. 39 (right column panels), and nicely follow the exponential behavior as predicted. In this Figure, the average on different experiments carried out at different pulling velocity are compared (or the native hairpin, similar data are obtained for the mutated ones), demonstrating that the coexistence force does not vary.

The coexistence force is obtained also in this case from the crossing point of the two straight $k(f)$ lines in the semi-log plots. The f_c data from pulling experiments on the 20bp hairpin are shown in Table 1 above; we note that the similar data from hopping experiments on the 20bp hairpin (also in the same Table) are affected by a much larger

error (the case of the GA mismatch is notably off, likely the relaxation of hairpin configuration was in this case incomplete).

By using the Eq. (2.23), (2.24), the free-energy difference can be extracted, and compared to the results from the hopping experiments. The force-pulling experiments on the shorter 10bp hairpins gave results that are difficult to interpret, since the vertical distance between the folding and unfolding trajectories is very small, and the two force diagrams are nearly overlapping so that it is difficult to clearly identify the folded and unfolded state. For the 20bp hairpin, on the other hand, this method gave quite clear results for the free-energy difference between the two states, reported in Table 3. Although in principle more reliable, because of the reasons discussed above, the values of ΔG^0 obtained by the force-pulling method are nevertheless not much different from those obtained in the hopping experiments (Tab. 2). By comparing the raw experimental data to the theoretical predictions of the NN model, a distinct underestimation of $\sim 3-9\%$ can be noticed, similarly to what observed for the results of hopping experiments above. This discrepancy will be considered further in the Discussion section below.

Table 3 also reports the values of the distance $x_{NU} = x_{N\to B} + x_{B\to U}$, the two terms being respectively deduced from the folding/unfolding kinetic rates.

As already noted, the absolute values of free-energy in Tab. 2 and 3 appear to be systematically lower than the nearest-neighbor (NN) model prediction.[30] In principle, this discrepancy could be due to several reasons, such as a systematic error in the force calibration, or some inadequacy of the additive NN model for the system considered. However, one important information comes from the accompanying molecular dynamics simulations reported in Chapter 5, that is the first base-pair attached to the dsDNA handles appears to be always opened, even at zero applied force. This same effect was observed also in the longer hairpins of the study by McCauley et al.[74] If the NN model is corrected by ignoring the contribution of the first G-C pair in both the 10- and 20bp hairpins, the values indicated as "NN corrected" in Tab. 3 and 4 are obtained, which show a much better agreement with our experimental data.

Table 3: Free-energy differences (kcal/mol) between the folded/unfolded state for the 20bp DNA hairpins from force-pulling experiments, with the NN model predictions. Data for the native configuration, and including a GA or GT mismatch defect. In the last column, the values of the x_{NU} at the coexistence force are also reported.

Hairpin	ΔG^0	Error	NN	NN corrected	x_{NU} (nm)
20bp native	28.1	2.4	32.05	29.52	16.6
20bp GA	26.9	3.1	28.45	25.92	17.0
20bp GT	25.0	1.9	28.14	25.61	16.3

Most importantly, the presence of a MM defect with respect to a perfectly matched native (WC) sequence can be traced, by looking both at variation of the coexistence force, reported in Table 1, and at the free-energy variation:

$$\Delta\Delta G = \Delta G_0^{WC} - \Delta G_0^{MM} \tag{3.18}$$

These latter are given in Table 4 for both sets of experiments. In general, the single-molecule force spectroscopy method is able to identify the presence of the mismatched base-pair. The value of $\Delta\Delta G$ is slightly lower for the G-A compared to the G-T, as also predicted by the NN model; however a considerable dispersion of the data is observed, also indicated by the rather large error bar. Note that, since we are taking differences, the values for the NN model do not change upon applying the above correction.

Table 4: Relative free-energy differences $\Delta\Delta G$ (kcal/mole) between the hairpin native (WC) and mismatched configuration from different experiments, with the NN model predictions.

experiment	hairpin	$\Delta\Delta G$	error	NN model
hopping	10bp GA	3.7	1.1	
hopping	20bp GA	2.2	1.2	3.6
pulling	20bp GA	1.2	3.2	
hopping	10bp GT	3.4	1.3	
hopping	20bp GT	4.3	1.4	3.9
pulling	20bp GT	3.1	3.0	

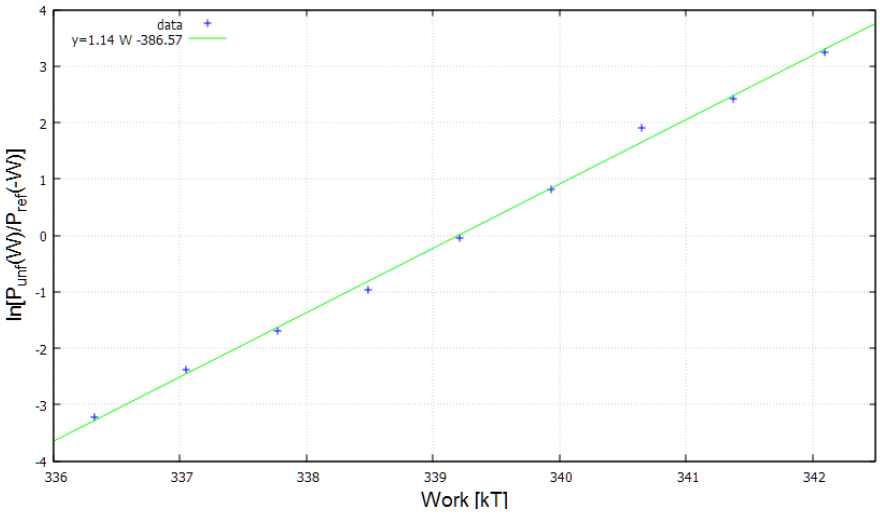


Figure 40: Verification of the Crooks Fluctuation Theorem for the 20bp hairpin with GA mismatch.

3.6.3 *An analysis by the Crooks' Theorem*

An alternative method to extract the free energy of the system out of equilibrium relies on the Crooks' Fluctuation Theorem (CFT),[34]. The CFT is based on the theory of non-equilibrium transformation between two fixed states. It relates the expected value for a functional of the the unfolding (U) and refolding (R) trajectories with work done on the transformation and with the free-energy difference between the initial and final states. In the special case where the functional is work (W) done along the trajectory gives a relation for the work probabilities distribution [77]:

$$P_U(W) = \exp[\beta(W - \Delta G)]P_R(-W) \tag{3.19}$$

The basic quantity extracted from the experiment is the work W_{NU} done during the unfolding ("forward" jump, $N \rightarrow U$) and the work W_{UN} during the folding (or "backward" jump, $N \leftarrow U$). The applicability of the CFT to our experiments can be tested by plotting the data of measured work along the different trajectories, as shown in Fig. 40 for the case of the 20bp hairpin with GA mismatch. The graph shows the logarithm of the ratio between the probability of observing a given value of W_{NU} and the probability of the same value of $W_{UN} = -W_{NU}$, as a function of W. As shown in the Eq.3.19, this kind of plot requires linearity between such log(ratio) and W, and we find this condition to be generally obeyed by our experimental data. However, the second condition is that when the work W is expressed in k_BT units, the value of the fitting coefficient must be ~ 1; while this is the case for the example shown in Fig. 40, we found this condition to be often violated in our experiments, therefore we are not presenting the corresponding free-energy data. We are investigating the possible origins of this discrepancy, since violations of the CFT may appear in the case of hidden contributions to the total work, which are not taken into account. Examples of such missing contributions could be,

e.g., the presence of a non-zero component of the force orthogonal to the pulling direction, or a torsional elastic term not considered in the WLC analysis.

3.7 FINAL SUMMARY

In conclusion in this part of the work, we used single-molecule force spectroscopy with optical tweezers, to demonstrate the ability of this experimental method to detect the small free-energy and force differences associated with a single, isolated MM in a DNA sequence. As a test system, we adopted a short self-complementary hairpin undergoing a reversible folding/unfolding transition, with the aim of extracting information about the differences induced in this dynamical transition by the presence of a G-A or G-T mismatch, inserted in the middle of the hairpin. This same system was used in a work earlier this year,[74] in which longer hairpin constructs including pairs of MM were used. With respect to that work, here we wanted to test the lower limits of the technique, to detect a single defect in the shortest possible sequence. To this end, we used hairpins of length 10 and 20 bp, in experiments carried out either at fixed displacement (*hopping*) or fixed applied force (*force-pulling*). Both methods showed pros and cons, namely:

(a) the hopping method is more adapted to the very short hairpins, since the folding/unfolding rate is inversely proportional to the length, therefore a large number of folding/unfolding transitions can be observed in this case, with a much better statistics;

(b) the force-pulling method is more adapted to the longer hairpins, since the difference between the folding/unfolding force path is proportional to the hairpin length, and for the shorter ones the two force branches are so close to be nearly indistinguishable.

In both sets of experiments, we could clearly demonstrate the detection of the MM in the hairpin, both as a net difference in the free-energy, ΔG_0, and as a shift in the coexistence force, f_c. The free energies were extracted in both cases by an analysis based on the Bell-Evans model [49],complemented by a Hidden-Markov model (HMM) analysis for the hopping experiments [161].

In the Bell-Evans model, the folding/unfolding process is described by a one-dimensional reaction coordinate performing a diffusive motion across a free energy potential barrier, reversibly separating two independent conformations of the molecule. The combined Bell-Evans and HMM analysis of experimental data showed that the folding/unfolding transition at such short hairpin lengths (10-20 bp) does not appear to imply the presence of intermediate states, between the two local minima of a fully-closed and fully-extended hairpin.

This was indirectly confirmed by MD simulations (described in the next Chapter 5) in which, despite a largely different time-scale at equilibrium and much faster deformation rates at non-equilibrium, no evidence for intermediate states was observed.

In conclusion, the results and potential implications of this study can be summarized by the following findings:

1. Single-molecule force spectroscopy by optical tweezers is capable of identifying the presence of a single mismatched base-pair in small DNA hairpins, as short as 10-20 base pairs. The mismatch affects the free-energy of binding and the coexistence force, at which the folded and unfolded configurations have the same probability of occurrence.

2. The folding/unfolding transition in short hairpins is properly described by a two-state model, without evidence for intermediate states, based on the analysis in terms of the Bell-Evans and Hidden-Markov model.

3. Molecular dynamics simulations of the hairpin unfolding transition under external force, and of equilibrium excitations, support this view and provide important clues for the analysis of experimental data; simulations also suggest a variable degree of cooperativity in base-pair opening during the forced unfolding.

The folding/unfolding transition in short hairpins is therefore a relevant test bed for studying defects in DNA. Further work should concentrate in refining the experimental set up, in order to arrive at a better quantitative characterization of point defects, and move towards the study of the role of signalization and repair proteins in defect dynamics.[148] Molecular simulations can play a relevant role in assisting and guiding the experimental analysis.

4

MOLECULAR DYNAMICS SIMULATIONS OF DNA

4.1 WHY USE COMPUTER SIMULATIONS TO STUDY BIOLOGY?

Biomolecules are complex atomic systems. Their biological functions in the cell life, and generally in all biological processes, should be explainable in terms of physical and chemical principles: ultimately, their macroscopic behavior is dictated by the microscopic interactions between the atoms and electrons they are made of.

The main limit of such a mechanistic approach to biology is the complexity of the involved molecules, ranging from proteins, to lipids, to nucleic acids. Each of these molecules are composed by hundreds or thousands of atoms and, in mathematical terms, they are described as *many-body systems* with a huge number of degrees of freedom, strictly coupled via a host of molecular interactions, covalent bonds, hydrogen bonds, electrostatic and dispersion forces. Even by neglecting the quantum-mechanical nature of the molecular forces, originating from the distribution of electron density that holds together the atoms, analytical solutions of the forces approximated by classical, Newtonian mechanics equations of motion, are obviously impossible. Computer simulations try to circumvent this problem by using numerical methods, in which molecules and atoms are replaced by model systems capturing the key characteristics of the real ones.

Computer simulations are powerful tools for the study of large many-body systems. In fact, despite the modern developments of analytical and experimental approaches, in many cases computer simulations are the only available techniques to study complex physical-chemical phenomena at the atomic scale. Such techniques are often used to support analytical modelling, or to drive towards the understanding of the experimentally observed phenomena. In particular, the method of Molecular Dynamics, introduced in the late '50s to confront rather theoretical questions in statistical mechanics and fundamental fluid physics, has today gained a vast popularity in materials science, biochemistry and biophysics.

The traditional numerical techniques for the simulations of many-body atomistic systems can be generally divided in two main approaches: *stochastic* and *deterministic*, respectively represented by the Monte Carlo (MC) and the Molecular Dynamics (MD) simulation methods.

- MC simulations are based on the exploration of the configurational space of a mechanical system through random test displacements of the degrees of freedom of the system. In advancing from one step to another, the total energy of the configuration is used as discriminant: if the energy of the new configuration is smaller then the initial one, the displacement is ac-

cepted; otherwise the displacement is rejected, or accepted with a probability that depends on the Boltzmann statistics. In this way, a sampling of the statistical ensembles can be obtained, and the *equilibrium* thermodynamic properties of the system are calculated as averages on the system configurations. A "kinetic" version of MC can be introduced, to study non-equilibrium phenomena, however the meaning of the time-scale in this case is entirely arbitrary.

- MD is based on the solution of Hamilton's equations of motion to compute the variation of positions and velocities of all the degrees of freedom in the system. This gives MD an advantage, in that the whole phase space is explored, and not only the configuration space as in MC. In this way, it is possible to obtain information also on the system dynamics *out of equilibrium*, the microscopic phenomena being followed in real time. The downside compared to MC is that instead of just energies, also the forces between atoms have to be computed, a much more costly procedure which covers typically ~90% of the simulation time.

With respect to equilibrium thermodynamic properties, however, these two methods are equivalent. Both require the introduction of a model to describe the interactions among atoms, which should reproduce at best the experimentally observed properties of the system. Since size and time considerations rule out a quantum mechanical calculation of the forces, as it is possible e.g. in the framework of density-functional theory for very small molecular systems, the interatomic potential model must be empirically defined by a more or less complex functional of the atomic coordinates, either in analytical form or by numerically tabulated functions. In the second part of the thesis document (which was actually developed in the first 1 and 1/2 year), we will be interested in the computer simulation of the microscopic modifications of the DNA structure, and of its dynamics due to the presence of defects. Therefore we decided to adopt the MD approach to develop our research. However, despite its importance for a deeper understanding of microscopic phenomena, MD cannot be the *Holy Grail* of molecular biology, because limitations of accessible time-scales and length-scales are eventually imposed, by technological limits of the computers, by the computational precision in the numerical solution of trajectories, and by the variable fidelity of description of the atomic interactions. Although such limits are continuously being pushed back by technological and algorithmic improvements, a correct approach would then ensure a continuous exchange of information between computer simulations and laboratory experiments.

In this Chapter we will refrain from a detailed description of the MD methods, whose technicalities can be found by now in a number of books (see e.g. [5, 54, 123]). After a very concise introduction on the quantum- and statistical mechanics motivations of the method, we will rather focus on just a few specific extensions of the standard MD, which will be relevant for the analysis of the simulations presented in the nextcoming three Chapters.

4.2 A BRIEF PRIMER ON MOLECULAR DYNAMICS

4.2.1 *On the quantum foundation of classical Molecular Dynamics*

The so-called "classical" Molecular Dynamics method uses classical (Newtonian, more generally Hamiltonian) mechanics to describe the time evolution of the virtual mechanical system representing the molecules of interest. Apparently, this is in contrast with the modern physical theory that requires *quantum mechanics* (QM) to correctly describe the dynamical behavior of a molecular system. However, under some conditions, the applicability of classical mechanics finds its justification in quantum principles. In QM the state of an isolated mechanical system of N point particles is represented by a wave function ψ defined in terms of the generalized coordinates $\{\mathbf{q}\} = (q_1, \cdots, q_N)$ and of the time t. This function evolves according to the Schrödinger equation:

$$i\hbar\partial_t\psi(\vec{q},t) = \widehat{H}\psi(\vec{q},t) \tag{4.1}$$

The $\widehat{H}$ is the quantum Hamiltonian operator:

$$\widehat{H} = \frac{1}{2}\sum_q \frac{\hat{P_q}^2}{m_q} + \hat{V} \tag{4.2}$$

with $\hat{V}$ the potential energy operator, $\hat{P}_q = -i\hbar\hat{\nabla}_q$ the momentum operator, and $\hbar$ the Planck constant divided by 2π. Physically, according to the generally accepted interpretation, the wave function contains all the possible information we could ask about the system. Mathematically, it is an element of the square-integrable function space $\mathcal{L}$, in particular for a system containing N particles, $\psi \in \mathcal{L}\left(\mathbb{R}^{3N}\right)$.

In a molecular system the total number of degrees of freedom is 3 x the number of electrons of each atom, plus 3 x the number of atomic nuclei. Even neglecting the quantum dynamics of the heavy nuclei, the problem in numerically solving Eq.(4.1) is that the computational overhead grows exponentially with the number of electronic degrees of freedom. This means that even for the smallest biomolecules the computational resources required are huge. To move beyond the limits of the Schrödinger equation, it is possible to simplify this many-body quantum problem with some approximations and theorems, which allow to replace quantum particles by classical particles, actually mathematical points with a mass.

The first approximation is a direct consequence of the large difference between the electron mass, $m_e = 9.109 \cdot 10^{-31}$kg, and the nuclear mass that even for the lightest atom, Hydrogen, is $m_H = m_{proton} = 1.672 \cdot 10^{-27}$kg $\sim 10^3 m_e$. This difference allows to separate the motion of the fast electronic degrees of freedom from those of the much heavier ions. This approximation, called *Born-Oppenheimer*, is mathematically expressed by the following factorization of the wavefunction into separate subspaces:

$$\psi = \psi_{electrons} \otimes \psi_{nuclear}$$

To describe the nuclear wave function, $\psi_{nuclear}$, we can use the *de Broglie hypothesis* stating that quantum effects are dominant at a length scale of the order of $\lambda = 2\pi\hbar/p$. The nucleus kinetic energy is related to the system temperature, for an ion of mass M at temperature $T \sim 300K$ (typical range temperature for biological experiments), it is $\lambda_M \sim \sqrt{\frac{2\pi\hbar^2}{Mk_BT}} < 10^{-10}$ m, where $k_B = 1.380649 \cdot 10^{-23} J/K$ is the Boltzmann constant. Therefore, quantum effects for the nuclei are confined well inside the atomic radius and, consequently, it is reasonable to describe nuclear dynamics at high temperature as the motion of a classical point-like particle coinciding with its center of mass.

The general picture is that of a slow inertial nuclear core, described by classical mechanics, that moves in a rapidly relaxing electronic density distribution. The initial function space for the N-nuclei wavefunction reduces to a classical phase space $\psi_{nuclear} \in \mathcal{L}\left(\mathbb{R}^{3N}\right) \rightarrow (\vec{q}, \vec{p}) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N}$, with $\vec{q}$ and $\vec{p}$ the classical particle coordinates and momenta. The total energy expression for this approximated system is a function of the ions kinetic energy, T_{ions}, and of the the electron binding energy:

$$E_{tot} = T_{ions} + \epsilon_0\left(\psi_{electrons}; \overrightarrow{q}_1, \cdots, \overrightarrow{q}_N\right), \tag{4.3}$$

In a classical mechanics approach, electrons instantly follow the nuclear dynamics, constantly frozen in the ground state of each new nuclear configuration. The effect of the electron density distribution is then represented by some customarily chosen empirical interatomic potential function, $V_{eff}(\{\vec{q}\})$, instead of the fully quantum-mechanical term $\epsilon_0(...)$ in the total energy.

Moving from the quantum formulation to the classical approach for atomic trajectories, the system dynamics evolves in the 6N-dimensional phase-space describing an isolated system of N classical particles, according to the Hamilton's equations of motion:

$$\begin{cases} \frac{dq_i(t)}{dt} = \frac{\partial H}{\partial p_i} \\ \\ \frac{dp_i(t)}{dt} = -\frac{\partial H}{\partial q_i} \end{cases} \tag{4.4}$$

where:

$$H = H(\vec{p}, \vec{q}) = \sum_i^N \frac{p_i^2}{2M_i} + V_{eff}(\vec{q}_1, \cdots, \vec{q}_N) \tag{4.5}$$

is now the classical Hamiltonian, M_i is the mass of the atom i (including its electrons), and V_{eff} is the classical *interatomic potential*. The explicit mathematical form of the interatomic effective potential is inspired from the structure and chemistry of the real system, and the involved parameters are fitted on a set of experimental data and quantum-chemistry simulations (see Appendix E.2).

4.2.2 *Observables from microscopic atomic trajectories*

In statistical mechanics, the macroscopic thermodynamic state of an atomic system is represented by a *statistical ensemble* containing all the

possible microscopic states consistent with the external constraints: for the example of an isolated system of N atoms, all the microscopic states (values of coordinates and momenta of its atoms) having the same total energy E and the same enclosing volume V. Since not all the microscopic states of the ensemble have the same probability of occurrence, the ensemble is also characterized by the corresponding distribution function $\rho(\Gamma)$, describing the probability distribution of each microscopic state $\Gamma = (\vec{q}, \vec{p})$.

The most commonly used statistical mechanics ensembles are those that can be assimilated to typical experimental conditions, in which some of the thermodynamics variables (temperature, pressure, volume, etc.) are controlled. Common examples are:

- *Microcanonical ensemble*, constant-{NVE}
  The thermodynamic constraints impose conservation of the particle number N, volume V and total energy E. Because there are no exchanges of energy with the environment, the system is said to be *isolated*. Assuming the postulate of equal a priori probability, all the microscopic states with the same energy E have the same probability. The corresponding distribution function is:

  $$\rho_{NVE}(p, q) \propto \delta(H(p, q) - E) \tag{4.6}$$

  The Dirac's $\delta(x)$ function filters out the states with energy equal to E, from the ensemble of all possible microstates. Therefore the system path in the phase space conserves energy by construction, since only microstates with the same energy E are allowed.

- *Canonical ensemble*, constant-{NVT}
  In this case, the constrained variables are the particle number N, volume V and temperature T. Exchanges of energy with the environment are allowed, typically representing a system in contact with a thermal bath at temperature T. The distribution function ρ_{NVT} is given in this case by the Boltzmann distribution:

  $$\rho_{NVT}(p, q) \propto \exp(-H(p, q)/k_B T) \tag{4.7}$$

  describing the fact that in this case only the *average* energy of the system is imposed (which can be shown to be $\langle E \rangle = \frac{3}{2} k_B T$); therefore, microstates with different energies can be "visited" by the system along its phase-space trajectory.

- *Isothermal-isobaric ensemble*, constant-{NPT}
  Now the particle number N, pressure P and temperature T are imposed. Likely, this is the closest approximation to chemical or biological laboratory experiments where the pressure is given by the environment (or a confinement reactor), and the temperature is fixed by the exchange of heat with a thermostat (or just the room). The distribution function is now given by:

  $$\rho_{NPT}(p, q) \propto \exp(-(H(p, q) + PV)/k_B T) \tag{4.8}$$

again a Boltzmann-like distribution, for which the average system *enthalpy* $H = U + PV$ is constant.

Once the macroscopic thermodynamic variables are fixed, and the corresponding statistical distribution function is set, any thermodynamic *observable* Q can be obtained by the statistical average of a suitable *dynamical variable* $\hat{Q}(\Gamma)$, whose values are distributed over the ensemble of allowed microstates Γ in the phase space:

$$Q =< \hat{Q} >= \int \hat{Q}(\Gamma)\rho(\Gamma)d\Gamma \tag{4.9}$$

For example, it is possible to define the following dynamical variables associated with common physical observables:

- *Configurational energy* $U \rightarrow \hat{U} = V_{eff}(q_1, \ldots, q_N)$
- *Kinetic energy* $K \rightarrow \hat{K} = \sum_i \frac{\vec{p}_i^2}{2m_i}$
- *Total energy* $E \rightarrow \hat{E} = \hat{K} + \hat{U}$
- *Temperature* $T \rightarrow \hat{T} = \frac{2}{3Nk_B}\hat{K}$
- *Pressure* $P \rightarrow \hat{P} = \frac{N}{V}k_B\hat{T} + \frac{1}{3V}\sum_i \vec{r}_i \cdot \vec{f}_i$

On the other hand, the observables measured in experiments can be seen as a time-averages of the corresponding dynamical variable along a given phase-space trajectoy $\Gamma(t) = (\vec{p}(t), \vec{q}(t))$:

$$Q = \bar{Q} = \lim_{T\to\infty} \frac{1}{T}\int_0^T \hat{Q}(\Gamma(t))dt \tag{4.10}$$

The *ergodic hypothesis* of statistical mechanics assumes that, for a system at the equilibrium, the ensemble average and the time average coincide:

$$Q = \bar{Q} =< \hat{Q} > \tag{4.11}$$

In other words, in the limit of infinite sampling time the system explores all the possible microscopic states with a frequency proportional to the statistical probability expressed by its statistical ensemble density ρ.

In this respect, a MD simulation tries to imitate the experimental situation, in that the observables are obtained from the simulated trajectory by means of the time averages of dynamical variables in Eq.(4.10), under the ergodic hypothesis. However, it is worth noting that MD simulations can be properly extended also to simulate non-equilibrium thermodynamics conditions.

4.2.3 *Performing the simulations*

The description of the behavior of complex molecules by MD simulations, in particular of biological polymers such as DNA, RNA or proteins, can address different levels of detail in the fictitious mechanical

structure that mathematically represents the molecule. The particular model used to perform a simulation depends on the quality and quantity of information that we want to obtain on system behaviour during the time evolution of the simulated experiment, and on the computational resources available.

The equations of motion obtained by the Hamiltonian in Eq.(4.5) describe the dynamics of an isolated system. On the other hand, biological processes do not take place in isolated systems but they are in contact with a thermal bath at physiological temperature, and are subject to the external pressure. In fact, MD simulations of any kind take place in a *simulation box*, mathematically defining the boundaries of the containing volume, which also identifies the possible interaction with the external constraints. The simulation box is made periodic in space by applying toroidal boundary conditions (see Appendix E.1), to represent a virtual infinitely extended system.

In order to apply these thermodynamic conditions in MD, different implementations of a numerical "thermostat" and "barostat" have been introduced, such as the Berendsen or the Nosé-Hoover formulations, each with slightly different effects on the system dynamics, but all generically acting via the "walls" of the periodic simulation box. A different method is based on the *Langevin dynamics*; notably, the Langevin scheme was already useful in the description of Kramer's reaction-rate theory (see Section 2.2.1; more detailed descriptions of thermostats and Langevin are given in the Appendix E.6).

The definition of the effective potential $V_{eff}(\vec{q}_1, ..., \vec{q}_N)$ acting between all atoms in the system is the central point in the MD description of a molecular system, and it should correctly reproduce the observed properties of the molecule. The sets of parameters describing the interatomic potential in MD are commonly called a *force field*, the interactions being customarily divided into *bonded* and *non-bonded* (see Appendix E.2).

Depending on their intended area of application, several force field schemes have been produced over time, with different parame-trizations for the molecular bonding and non-bonding interactions. Some of the most commonly used ones for all-atom simulations of nucleic acids and proteins are AMBER ("Assisted Model Building with Energy Refinement", developed by P. Kollman and collaborators at UCSF), GROMOS ("Groningen Molecular Simulator", jointly developed by H. Berendsen in Groningen and W. Van Gunsteren at ETZH), and CHARMM ("Chemistry at Harvard Macromolecular Mechanics", developed by M. Karplus and collaborators at Harvard). In the present work, we adopted the latter, in the version CHARMM-27 [98, 99] and its extensions to treat nucleic acids [51, 97].

After the start of this thesis, a much improved version of the AMBER force field has been developed by the group of M. Orozco in Barcelona (labelled *parmbsc*1, [73]), especially designed to suppress some problems arising in extended-time DNA simulations. Nevertheless, strict comparisons between CHARMM-27 and existing nucleic-acid AMBER force fields (*parmbsc*0, [116] A.D. 2007) have been carried

out [124, 146], warranting that the results of long-time, finite- temperature molecular simulations of nucleic-acid fragments with largely different configurations are internally consistent, and are able to correctly reproduce the key structural quantities (bond angles, hydrogen-bond structure, base tilt, twist, shuffle, etc.) when compared to experimental data. This, however, does not exempt from taking care, in all cases, of performing long preparatory annealing cycles of the water and ion background, while keeping the DNA still in the simulation box, to allow a realistic arrangement of the counter-ions around the phosphate backbone, prior to starting the actual production runs.

For the very-large-scale simulations that are possible with massively parallel supercomputers, the subroutines that compute the force acting on each particle are so important that they are written in a way compatible with the detailed computer architecture. To perform our MD simulations we used both the NAMD [120] and GROMACS computer codes [18, 92], on the large national supercomputing facilities made available by CNRS (`http://www.idris.fr`) and French Ministry of Research (`http://www.cines.fr`).

4.3 ANALYSIS OF MD TRAJECTORIES

The very detailed knowledge of atomic trajectories (that is, the time sequence of coordinates and momenta of all the atoms in the system) as made available from MD simulations, allows for countless different analyses of the physical (and in part, chemical) behavior of the simulated system. Ordinary equilibrium thermodynamic properties, such as temperature or pressure, can be calculated as explained in the previous Section. Since MD is performed at finite temperature, further quantities can be deduced by looking at the RMS fluctuation of the basic thermodynamic variables, such as specific heats (energy or enthalpy fluctuation), thermal expansion coefficients (volume fluctuation at constant temperature), compressibility (volume fluctuation at constant pressure), and so on.

Moreover, an much richer information about the transport properties of the system can be obtained, by calculating *correlation functions* between any two observables A and B via the celebrated Green-Kubo formula (derived from the fluctuation-dissipation theorem [60, 83] and linear response theory):

$$C_{AB}(\tau) = \langle \hat{A}(t) \cdot \hat{B}(t+\tau) \rangle \tag{4.12}$$

(including the "autocorrelation", i.e. $A = B$). For example, the Fourier transform of the velocity-velocity autocorrelation function:

$$C_{vv}(\tau) = \frac{1}{N} \sum_{i=1}^{N} \langle \vec{v}_i(t) \cdot \vec{v}_i(t+\tau) \rangle \tag{4.13}$$

provides the vibrational spectrum of the system; or the density-density autocorrelation function:

$$C_{\rho\rho}(\vec{r}) = \frac{1}{N} \sum_{i=1}^{N} \langle \rho_i(\vec{r}_i) \cdot \rho_i(\vec{r}_i + \vec{r}) \rangle \tag{4.14}$$

whose Fourier transform is related to the x-ray scattering structure factor $S(\vec{k})$; or again, the heat-flux autocorrelation function, defining the thermal conductivity κ:

$$C_{jj}(\vec{r}) = \frac{1}{N}\sum_{i=1}^{N}\langle J_i(t)\cdot j_i(t+\tau)\rangle = (3k_B T^2 V)\ \kappa(\vec{r}) \tag{4.15}$$

In the present work, autocorrelation functions have been calculated to obtain the vibrational spectra and principal-component analysis, the results being reported in the following Chapters 6 and 7.

Other powerful analyses of the MD equilibrium trajectories will be described in some detail in the remainder of this Section. As said in the general summary, MD also allows the simulation of non-equilibrium process, among which notably a virtual version of the force-pulling experiments described in the preceding Chapter. Even if not directly comparable to the experimental situation, because of time-scale limitations, such an analysis could provide relevant information on phenomena not directly accessible in laboratory. In Section 4.3.4 below, we will describe the *steered molecular dynamics* (SMD), as the numerical counterpart of the single-molecule force spectroscopy experiments.

4.3.1 *Study of the molecular vibrational modes*

As already described, MD simulations can rapidly become computationally expensive, typically sampling atomic motions occurring at very best on the microsecond timescale. MD is therefore not particularly suited to sample conformational motions that occur on much longer timescales (typical example, the DNA hairpin unfolding transition under external force).

Looking at the vibrational spectrum of molecules is a useful complementary technique to unravel their collective modes, thus providing support in understanding the functional mechanism due to the (only apparently) random fluctuations of macromolecules. Vibrational analysis describes all possible deformations that a molecule, modeled as an ensemble of decoupled harmonic oscillators, can undergo around a stable equilibrium configuration. Under physiological conditions, molecules do not always have the possibility to explore all the available conformational space; more often, they explore only a narrow region around the local energy minima, since the eventual transition to one misfolded or improperly folded state could inhibit the entire molecule functionality in the living organism [125]. This restricted fluctuation region defines a *native state*, composed by a sub-ensemble of micro states that share a common secondary structure.

If we consider an ensemble of P closely similar structures, defined by the set of their coordinates $\{\vec{r}_i\}_{k\in P}$, the vibrational analysis is an orthogonal linear transformation, from the standard coordinate system representing the $\{\vec{r}_i\}$ in $\mathbb{R}^{3N}$, into a new reference frame of *collective coordinates* identifying the dominant directions of structural changes.

Vibrational analysis aims at detecting those small-wavelength collective modes accessible to biomolecules, so as to track down the large-scale physiological movements of a specific stable configuration. Comparing the result of this analysis between the *native* form of a biomolecule and its *mutated* version could reveal key differences in the degrees of freedom responsible for reducing the functionality.

Two main methods of vibrational analysis have been developed in the field of biomolecular simulations:

- Normal Mode Analysis (NMA)
- Essential Dynamics (ED)

These techniques will be succinctly described in the following two subsections.

4.3.2 *Normal Mode Analysis*

Normal-mode analysis (NMA) is a technique to investigate the vibrational motions of a mechanical system around a local minimum, and could be used to study structural deformations and rearrangements of a molecule. The underlying assumption is that at the equilibrium the system fluctuates around a single conformation and that the nature of these thermally-induced fluctuations can be calculated assuming an harmonic potential about the equilibrium position of the atoms, so the variations are confined to the neighborhood of a minimum [14, 140].

If we consider an ensemble of N particles in a Cartesian reference frame, described by 3N coordinates $\{q_i\}_{i=1...3N}$ about a local minimum, $\{q_0\}$, it is possible to approximate the potential energy V with the first terms of its Taylor expansion:

$$\begin{aligned} V(\{q\}) &= V(\{q_0\}) + \frac{1}{2}\sum_{i,j}(q_i - q_{0i})\frac{\partial^2 V}{\partial q_i \partial q_j}(q_0)(q_j - q_{0j}) + \dots \\ &\approx V(\{q_0\}) + \frac{1}{2}\sum_{i,j}(q_i - q_{0i})V''_{ij}(q_0)(q_j - q_{0j}) \end{aligned} \tag{4.16}$$

where the first-order linear term is null by symmetry about a locally-harmonic minimum. By writing the equations of motion, we obtain the following 3N-equation coupled differential system:

$$M_{ij}\frac{d^2 q_j}{dt^2} = V''_{ij}(q_0)(q_j - q_{0j}), \quad i = 1, ..., 3N \tag{4.17}$$

or:

$$\frac{d^2 q_k}{dt^2} = m_i^{-1}\delta_{ki}V''_{ij}(q_0)(q_j - q_{0j}) \tag{4.18}$$

where $M_{ij} = m_i\delta_{ij}$ is the diagonal matrix of particle masses. Note that also $M_{ij}^{-1} = m_i^{-1}\delta_{ij}$ is positive definite and diagonal, whereas the matrix of second derivatives (or *Hessian*) V''_{ij} is symmetric, therefore the RHS product is, in turn, a symmetric $(3N \times 3N)$ matrix that can be diagonalized, using an orthogonal transformation T_{ij}.

By applying the same orthogonal transformation to the particle coordinates:

$$T_{ij} q_j = x_j \tag{4.19}$$

a whole new set $\{x_i\}$ is obtained, in which all the original coordinates are mixed in each x_i. These new coordinates describe *collective motions* of all the particles at the same time. The Newtonian equations of motion for the transformed set are easily obtained:

$$\frac{d^2 x_i}{dt^2} = (Tq)_i = (TM^{-1}V''T^T)_{ij}(Tq)_j = \omega_j^2 x_i \tag{4.20}$$

The eigenvalues ω_j^2 are all positive, since in a local minimum each displacement from the rest position gives an increase in the potential energy. The equations (4.20) have harmonic solutions, and the eigenvalues are the *vibrational frequencies* of the collectives modes of all particles about the minimum, the corresponding eigenvectors describing the concerted motion of all the particles in each mode. Note however that they do not give the absolute value of the displacement amplitudes. These latter have a temperature-dependent magnitude $|\Delta q_i| \sim \sqrt{k_B T/\omega_i}$, meaning that the modes with lower frequency, or smaller wavelength, have the larger spatial oscillations. Because they correspond to smaller excitation energy, these "soft" modes are as well the most easily accessible to the N-particle system. Several studies have shown that these low-frequency vibrational modes correspond to functionally relevant motions in proteins, and that conformational transitions follow one or a few normal modes.

The NMA technique is capable of extracting the fundamental vibrational states of the molecular structure, even if the molecule could be deformed into less probable microscopic configurations. Experimental techniques as the NMR and X-ray diffraction methods are capable of directly accessing the vibrational spectrum of a molecule; their importance in this respect cannot be overstated, as they provide essential experimental information about the molecular structure. The NMA provides a direct comparison with the experimental data, allowing to compare vibrational frequencies and relative mode amplitudes.

Anyway, the NMA method has also important limitations, because of the assumption that the system can be described by a parabolic potential, therefore for large vibrational amplitudes the resulting configuration could be but a raw approximation of the real system vibrational state. Moreover, from the computational point of view diagonalization of a $(3N \times 3N)$ matrix may become very hard with increasing N, since the best numerical methods have a scaling of order $\mathcal{O}(N^3)$.

4.3.3 *Essential Dynamics*

Essential dynamics (ED), also called "principal component analysis" (PCA), is a method aimed at extracting the most relevant collective motions of the atoms during a MD simulation [7, 38]. The objective

of this method is to separate the configurational space into two subspaces: an *essential* subspace, containing just a few degrees of freedom relative to the ample, anharmonic movements that dominate the global molecular motion; and the remaining subspace, where the degrees of freedom are considered physically constrained, and contains mostly harmonic vibrations. The two subspaces are obtained from the unique diagonalization of the covariance matrix of the atomic displacements (actually, a cross-correlation matrix according to Eq.(4.12) above) for the set of N atoms:

$$\mathbf{C}_{ij} = \langle [\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle] \cdot \left[\mathbf{r}_j(t) - \langle \mathbf{r}_j \rangle\right]^T \rangle \tag{4.21}$$

where $\mathbf{r}_i(t)$ is the time trajectory of each atom i, and $\langle ... \rangle = \frac{1}{\tau}\int_0^\tau ...dt$ indicates time-averaging over the trajectory of length τ. By construction, the matrix $\mathbf{C}_{ij}$ is symmetric of rank 3N and sparse, because of the short range of the interatomic forces which make only the close neighbors j of each atom i to be spatially correlated, thus having non-zero matrix elements. Therefore it can be diagonalised by standard methods as:

$$\mathbf{C} = \mathbf{V}\mathbf{U}\mathbf{V}^T \tag{4.22}$$

with $\mathbf{U} = \langle \mathbf{q}\mathbf{q}^T \rangle$ the diagonal matrix of the eigenvalues u_i, $\mathbf{q} = \mathbf{V}^T(\mathbf{r} - \langle \mathbf{r} \rangle)$ is the new set of coordinates after orthogonalization, and $\mathbf{V}$ the solution matrix whose columns are the eigenvectors $\mathbf{v}_i$, $i \in 3N$. The eigenvectors can be ordered according to the amplitude of the corresponding eigenvalue. Quite surprisingly, it is generally found that only the first few eigenvalues (Figure 41) of the ED represent the largest part (i.e, have the largest weight) of the total atomic displacements.

Similarly to the subdivision of the eigenvector space, also the new coordinates $\mathbf{q}$ can be separated into essential and non-essential. Let us indicate $\{\xi_i\}_{i\in M}$ the principal eigenvalue subset, and $\{s_j\}_{j\in 3N-M}$ the set of eigenvalues in the remaining subspace (with $M \ll 3N$). These latter can be shown to have a Gaussian narrow distribution with mean value zero, and behave as harmonic oscillations with a large force constant, which can be treated as mechanical constraints [7]. It is then possible to approximate the molecular mechanics in the essential subspace by setting the harmonic oscillation degrees of freedom to zero (as if their force constants were actually infinite), and rewrite the effective potential describing the essential dynamics of the molecule in the approximated form:

$$V(\mathbf{q}) = V(\xi, s) \simeq V(\xi; s = 0) + \frac{1}{2}\sum_i k_i s_i^2, \qquad i = M+1, 3N-6 \tag{4.23}$$

(the 6 degrees of freedom corresponding to rotation and translation of the center of mass have been excluded from the sum). As a consequence, the most important part of the molecular displacements are contained in this approximate representation of very low dimensionality, compared to the complete description of the molecular dynamics.

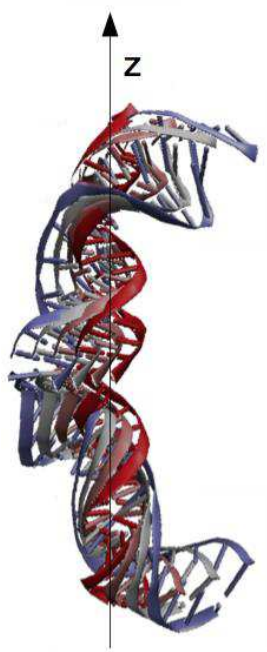


Figure 41: Action of the first eigenvector on the initial DNA configuration for the model of double strand break with two base pairs of distance. Frames are colored from red to blue, corresponding to increasing values of the virtual displacement given by the eigenvector (note that *this is not* a representation of the actual displacement path in time, but only an ordered projection of one of its main components). It can be seen that this collective mode corresponds to a strong bending of the molecule about an axis perpendicular to the helical axis of symmetry ($\hat{z}$ in the figure).

In comparison of NMA, the ED (or PCA) method seems to provide a more direct description of the large-scale conformal transitions, because is not confined to the harmonic motions about the local energy minimum. Therefore, it should be more accurate in considering the complex landscape of the potential energy surface in biological molecule. However, the Achille's heel of the method is in the requirement of computing a very long trajectory, from which the covariance matrix is extracted: if the trajectory is not long enough to sample the large-scale motions of the molecule, the ED method falls back to the NMA. When the native state of the molecule is characterized by a deep potential minimum (as it is often the case), it may be very hard for a straight MD simulation at room temperature to pull it out of its ground state, in order to explore additional conformations. Therefore, special "acceleration" methods must be devised, in order to facilitate the conformational transitions, such as the "steered-MD" method described in the next Section.

From the computational point of view, the ED anaysis can be more efficient than NMA, since very efficient diagonalization methods exist for sparse matrices like the covariance matrix; furthermore, if one is interested only in the first few eigenvalues (usually the first 4 or 5 suffice), iterative diagonalization methods such as the Lanczos can be used, with a much better scaling of $\mathcal{O}(N^2)$.

In our analysis on DNA defect dynamics described in Chapters 6 and 7 we used both the NMA and ED techniques, in the search to identify characteristic signatures of the defects. Notably, by ED we could track the first larger eigenvalues and the subspace they represent, to observe which deformation they describe; by NMA we compared the vibrational spectrum to experimental infrared and Raman spectra of DNA, trying to identify possible peak shifts induced by the presence of defects.

4.3.4 *Steered Molecular Dynamics*

The standard MD methods described earlier in this Chapter find their main limitation in the time-scale of the events simulated, which as said can reach in the best case the microsecond time scale. Moreover, such exceedingly long simulations require a huge amount of computer time, making them a "one-of-a-kind" event, and hampering the repeat of the simulation with many different initial conditions, necessary to accumulate sufficient statistics on the possible conformations explored by the molecule. To address this weakness, a number of "accelerated sampling" methods have been developed over the years, all which focus on enhancing the probability of *rare events*, i.e. those that would occur with negligible probability over the MD time scale at standard temperature and pressure. "Hyper-" [165], "Meta-" [86], "Replica-" [164], "Temperature-accelerated-" [143], "Learn-on-the-fly-" [35] versions of MD have been introduced, as well as MD-based variants such as "nudged-elastic bands" [76], "blue-moon ensemble" [147], "reactive flux" [13], "activation-relaxation" [16], and more, in a seemingly endless zoology of increasingly complex formulations, each one with advantages and disadvantages according to the specific application.

As described in the preceding Chapters, single-molecule force spectroscopy are paramount tools for the study of many properties of biomolecules, including conformational transitions. Among the rare-evens acceleration methods, *steered molecular dynamics* (SMD) seeks to complement these observations and provide atomic level descriptions of the underlying events (see e.g. [61, 71, 72]). Closely imitating single-molecule experiments, SMD applies external force vectors to an ordinary MD simulation, to manipulate biomolecules in order to probe mechanical functions, as well as to accelerate processes that are otherwise too slow to model. Therefore, the SMD appears as an ideal choice to pursue in the development of the computer simulations in this thesis.

The external force in SMD is applied on a chosen set of atoms in the molecular system, according to one of these pulling protocols:

- *Constant pulling speed*: a chosen set of particles (for example, the terminal base-pair of a dsDNA fragment) is bound with an harmonic potential U_0 to a virtual atom (a point) moved at constant velocity v_0. Another set of particles is fixed at zero displace-

ment, to provide the reaction. The force applied to the molecule depends on the extension of the virtual spring:

$$\vec{f} = -\vec{\nabla}U_0 \tag{4.24}$$

$$U = \frac{1}{2}k_0\left[v_0 t - (\vec{r}-\vec{r}_0)\cdot\hat{n}\right]^2 \tag{4.25}$$

as measured by the displacement $(\vec{r}-\vec{r}_0)$ of the virtual atom, projected along the direction $\hat{n}$ of the applied force.

- *Constant pulling force*: a constant force is applied to the set of moving particles, with no need for an intermediate virtual spring. Usually in this case the force vector is applied to the position identified as the center of mass of the chosen set of particles.

The succession of configurations drives the molecule along a particular conformation change, solicited by the external force: for example, the unfolding path of the DNA hairpin, or the breaking apart of a damaged fragment of dsDNA. From this force-pulling simulation we can calculate the free energy profile of the energy barriers that characterize the conformational transition of the molecule of interest.

The potential of mean force (PMF, [69]) is a method to extract the free energy difference ΔG from a sequence of atomic configurations, biased along the reaction coordinate λ that brings the system from the initial state N to a final state U, by estimating the force f_λ necessary to quasi-statically hold the system at each different value of λ:

$$\frac{\partial}{\partial\lambda}\Delta G_{N\to U} = \langle f_\lambda\rangle \tag{4.26}$$

where $\langle ...\rangle$ in this case means averaging over $N < \lambda < U$. Since the reaction coordinate is arbitrarily chosen, additional care must be taken to allow the system to "explore" as much as possible the nearby configurations, to increase the statistical sampling of possible intermediate states between N and U. This can be achieved by the so-called "umbrella sampling" technique [159]. To obtain the PMF by umbrella sampling at discrete values of λ, the original ("true") molecular potential is biased by an additional harmonic potential $V'(\lambda)$ at each point along the reaction coordinate; this allows the system to sample configurations in a small parabolic well around each λ. The probability of finding the system at λ is now biased, $P'(\lambda)$, and the unbiased estimate of G is:

$$G(\lambda) = -k_B T \ln P'(\lambda) - V'(\lambda) + c \tag{4.27}$$

with c an undetermined constant that disappears when computing free-energy differences ΔG bewteen any two states. Finally, the discrete values of $G(\lambda)$ between N and U must be smoothly connected. Usually the "weighted-histogram" method is used for this purpose [67, 84]. In practical calculations, one extracts ~100-200 configurations from a force-pulling simulation, spaced by typically 50 ps along 5-10 ns of trajectory. Each configuration must be equilibrated for a few more ns at under constant-{NVT} conditions, while biased with the

harmonic "umbrella" potential of variable strength, progressively reduced to zero to obtain the unbiased limit.

Eventually, a large enough force, or displacement velocity, will get any free-energy barrier to be overcome. However, for too fast deformation rates the sampling of the phase space becomes irrelevant, and the useful information on the transition states is lost. It is like shooting a cannon ball across a mountain range: the ball goes indeed from the initial to the final point, but all the important details of the landscape are obliterated. Fully-atomistic MD simulations have proven very useful in describing details of the response to mechanical forces [171], also indicating the occurrence of structural transformations of DNA [94, 132], as well providing insight into exotic structures such as the i-motif and the G-quadruplex [80, 139]. However, attaining a one-to-one comparison with single-molecule experiments is complicated, chiefly because the time-scales (and therefore the deformation rates) accessible to MD are far from comparable with the experimental ones. The best SMD simulation can achieve speeds of ~0.1 cm/s, orders of magnitude larger than even the faster AFM. On the other hand, single-molecule techniques for stretching DNA of contour length less than a few hundred kb, such as those that can be simulated by MD, are affected by various experimental difficulties. The study of individual processes (i.e., the landscape details across the mountain range) requires the ability to isolate the event in a relatively small molecule, in order to have a good signal-to-noise ratio. This is true not just for the identification of point defects (mismatches, SSB and DSB) of interest in this work, but for many other interesting biological events, e.g. histone binding, or protein-mediated looping of DNA, occurring over wide length- and time scales. The kind of theory-experiment comparison that we will present in the next Chapter on the DNA hairpin simulations is rather unique, thanks to the capability of the Barcelona team to work with extremely small molecules, which MD can simulate on a 1-to-1 length scale, albeit not on the same time-scale.

4.4 DNA IN MOLECULAR DYNAMICS SIMULATION

There are some assumptions that must be verified to justify the description of the DNA molecule with classical MD. We discussed in Section 4.2.1 how the quantum nature of molecular bonds can be described by an empirical effective potential, provided that the nature of the chemical bonds does not change during the simulation. This means that, within the framework of classical MD, it is not possible to directly study the process of radiation interaction and formation of the damage caused, e.g., by free radicals. To this aim, a combined quantum-classical simulation (or QM/MM) may be used [46, 56, 145, 171], in which a very restricted region of the molecular system is completely described by a quantum wave function, whereas the rest of the molecule is described by classical MD, and the contact between the two zones is assured by some kind of intermediate region. This type of simulations are extremely costly in terms of computational

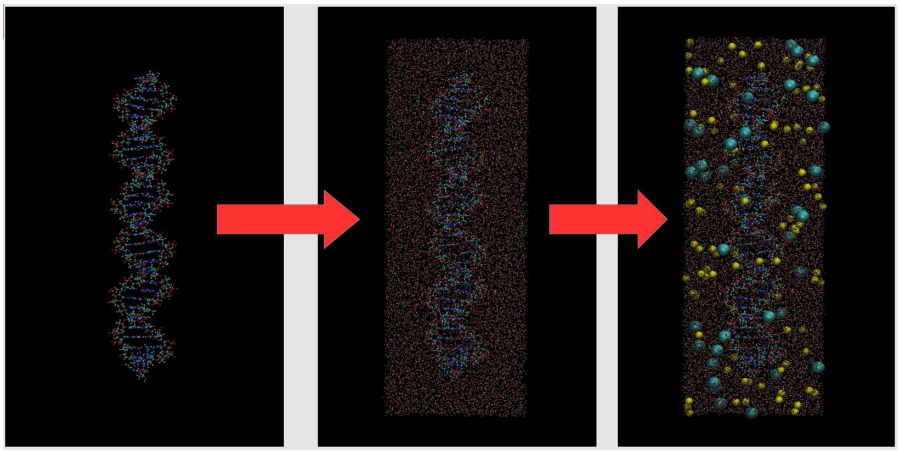


Figure 42: Preparation of the initial configuration for a linker dsDNA. The initial double-helix raw structure (left) is solvated in a large water box (center), finally Na^+ (yellow beads) and Cl^- (blue beads) ions are added, to obtain charge neutrality plus a fixed concentration (right).

resources, due to the necessity of solving the quantum-mechanical equations for the inner system, and propagate the effects of the changing electron distribution on the nuclei and the atoms in the intermediate region.

Therefore, we decided to resort to a more restricted study of the behavior of DNA samples with already formed defects, and the main interest was focused on the mechanical properties that characterize the subsequent evolution of the defects. Of course, one must be aware of the approximations implicit in any reduced empirical scheme, compared to quantum-chemical methods, and exercise caution when comparing simulations to experimental data (see e.g. [39] and references therein). The risk is however less critical if the simulations are used, as in the present work, to investigate general trends, as a guide in designing an experimental strategy, and avoiding the temptation of going too "chemical" in the inference, e.g. looking for fine details in the sequence-dependence of the results. Some known difficulties of empirical force fields for DNA reside, for example, in a possible overestimate of the stacking interactions whose origin may be difficult to trace back to a specific flaw of the potential: incorrect balance of hydrophobic/hydrophilic interactions, a poor electrostatic model for nucleosides, incorrect van der Waals terms for nucleobases, have all been indicated as possible origins, however without conclusive results. Another well-known issue is the neglect of a specific polarization term in current classical simulations. While this is a major source of uncertainty, the community has been very reluctant to use polarized force-fields, not only because of the considerable additional computing cost, but also because the final results were not exciting. Notably, the CHARMM community has recently released an efficient polarization algorithm based on Drude's oscillators [134], which appears to provide a good representation of the DNA duplex in the μs-regime, albeit at the price of a considerable extra computational cost.

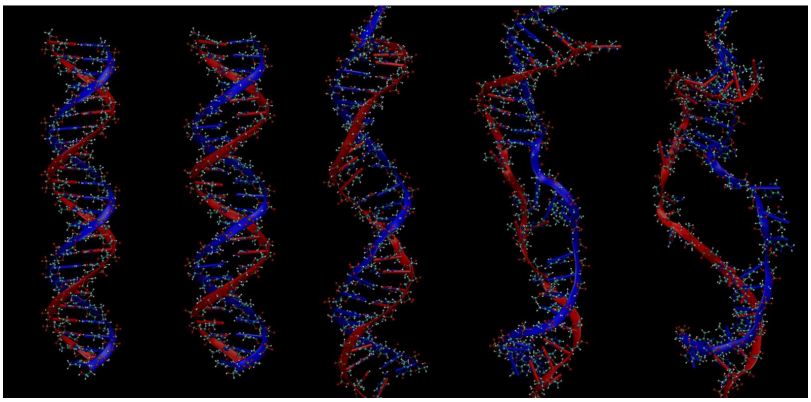


Figure 43: Simulation of a DNA fragment with implicit solvent performed with NAMD. In the figure from left to right, a succession of the polymer configurations at 1-2 ps intervals. It is observed that the two strand rapidly denaturate and the double-helix shape is lost.

In physiological conditions DNA is dissolved in a solution of water and ions. Such elements are fundamental for DNA structure stability because dehydration and changes in salinity could drastically modify the molecular shape (denaturation) and its binding with other biomolecules, notably the histones. In general, the system should be charge-neutral to avoid problems in the computation of electrostatic potential (the conditionally convergent Madelung sum of point charges). In computer simulations of DNA, and biomolecules in general, one tries to recreate conditions similar to those found in the living organism, so the simulation box is filled by water molecules, and some concentration of ions (most often Na^+ and Cl^-) is added, both to neutralize the DNA charge from the negative phosphate groups, and to reach a standard physiological concentration of ~ 0.15 M (Figure 42).

The number of water molecules in a large simulation box can grow very fast, the simulation time spent in calculating water interactions becoming overwhelming compared to the time dedicated to DNA. One possible solution is to use a simplified algorithm, in which the solvent is described in an implicit way. This method introduces new terms in the equation of motion to simulate the effect of a solvent as a continuous background, with the advantage of reduce the number of particles in the system because no water nor solvent are added in the simulation box. Unfortunately, this algorithm does not ensure the stability of DNA double helix, as shown in Figure 43. Therefore, for long-time MD simulations it cannot be avoided to treat individual solvent molecules explicitly.

Despite a simple molecular structure, the H_2O molecule of water presents more than one challenge to MD simulations. The electronegative oxygen forms two covalent bonds with the hydrogens, the geometrically defined angle H-O-H being of 104.5Å. The asymmetric electron charge distribution gives water important polar properties, a permanent dipole moment of 1.87 D, and an unusually large quadrupole of ~4-6 D·Å in the plane perpendicular to the symmetry axis. The two lone pairs plus the oxygen ion make it possible to form up to four hy-

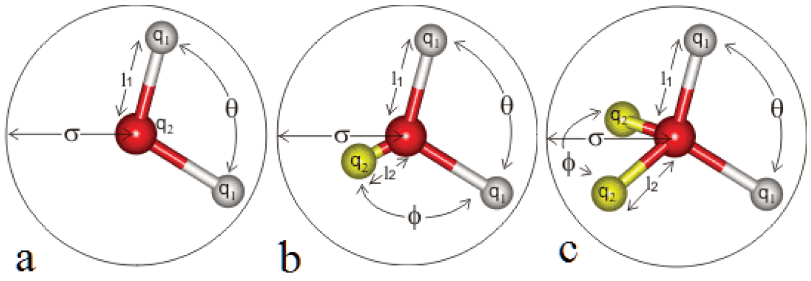


Figure 44: Water models in molecular dynamics simulation: (a) 3-site model, (b) 4-site model (c) 5-site.

Table 5: Different parametrizations for the two water models TIP3P and SPC/E. The parameters are referred to the Figure 44, in particular σ and ϵ are the parameter used in the Lennard-Jones potential to describe non-Coulombic long-range interactions.

Model	σ[Å]	ϵ [kJ/mol]	l_1 [Å]	q_1 [e]	q_2 [e]	θ [Â°]
TIP3P	3.15061	0.6364	0.9572	+0.4170	-0.8340	104.52
SPC/E	3.166	0.650	1.0000	+0.4238	-0.8476	109.47

drogen bonds, thereby leading to a tetrahedral coordination in both the solid and liquid phase.

These properties are described in classical MD by different models. Bulk properties of liquid water in MD simulations are affected, for example, by the system size, the method used for truncating long-range interactions, and the method used for temperature control. The 3-site model TIP3P uses three point charges to describe the molecule with effective charge assigned to the atoms, bond potentials to fix the H-O distance and H-O-H angle, and a long range potential (Lennard-Jones) to simulate the intermolecular interaction. More complex models could include 4- and 5-sites where fictitious particles are added to refine the polarizability properties of the molecule (Figure 44). Anyway, in our MD simulations this level of detail was not deemed necessary, especially because it comes with an increase of the number of degrees of freedom in the system, and of computational power accordingly. Therefore we used the standard 3-site models as the TIP3P or the SPC/E. The SPC/E water model is known to give the best bulk water dynamic properties and structure factors, whereas the TIP3P (both the original, and the CHARMM modified version) gives less structure at the level of second-neighbors shell, and faster dynamics with a diffusion coefficient of 5.6 (units of 10^{-9} cm^2/s), compared to 2.8 for SPC/E, and the experimental value of 2.3 for liquid water. However important they may be for fundamental studies of water properties, such differences should not impact in a significative way on the results of our simulations, since water provides only the background and it is often ignored in the analysis of DNA response.

4.5 WORK PROGRAM OF THE SIMULATIONS

In the following three Chapters, we will initially present the work on the DNA hairpin model reproducing the 10 bp native hairpin experimental arrangement. The goal of this MD simulation study was to verify some assumptions done in the polymer models used to interpret the experiments, and detect microscopical phenomena that could determine the discrepancy between the expected theoretical values for the free-energy of formation and the measured ones.

Subsequently, we present the study on the effects of different SSB and DSB damage to the backbone, in dsDNA. We have built a reference dsDNA fragment of 31 bp in B-form, restrained at its ends by elastic springs to be representative of the freely-exposed linker-DNA between two nucleosomes in the chromatin fiber. The DNA sequence has been randomly chosen, because at this stage we were not focusing on any particular sequence-dependent pattern. By cutting one or two bonds at a phosphate group in the backbone, we simulated the presence of a single-strand break, and three different type of double-strands breaks. This study will elucidate a number of structural and dynamic consequences of the SSB and DSB, and the way such defects may lead, in some cases but not always, to the breaking apart of the DNA fragment.

Finally, by focusing on the defect that showed the most relevant variation of mechanical properties, namely the DSB with only 1 base-pair spacing between the cuts, we have studied the damage mechanical response of the nucleosome. We performed microsecond-long MD simulations of nucleosomes including the DSB at various sites, to characterize the early stages of the evolution of this DNA lesion. The damaged structures are studied by the essential dynamics of DNA and histones, and compared to the intact nucleosome, thus exposing key features of the interactions. It will be shown that DSBs generally tend to remain compact, with only the terminal bases interacting with histone proteins. Umbrella-sampling calculations show that broken DNA ends at the DSB must overcome a free-energy barrier to detach from the nucleosome core. Furthermore, by calculating the covariant mechanical stress with a recently published formulation, we demonstrate that the coupled bending and torsional stress can be responsible for forcing the DSB free-ends to open up straight from the nucleosome body, thus making them accessible to damage-signaling proteins.

5

DNA HAIRPIN SIMULATIONS

After completing the experimental characterization of mismatch defects in DNA hairpins by force spectroscopy, and while working on the data anaysis and interpretation, it was all too natural to draw the attention to an accompanying theoretical modelling of the same systems. However, the time left before the end of the doctoral contract was not enough to allow for a complete study, therefore we had to make some drastic choices. In particular, given the extremely large difference between the experimental and MD time-scales, which make for exceedingly fast pulling rates and vanishing hopping rates in the simulations, a direct comparison between the theoretical and experimental results of force-pulling runs would not be possible. Therefore, we focused the simulations only on the native hairpin sequence, and ignored the hairpin with mismatches, aiming to provide additional insight at least on some molecular-scale phenomena that could be relevant to analyze and interpret the experimental results of Chapter 3. A more extended study would have been very desirable, notably as far as estimates of defect energies are concerned. However, the time constraints prevailed: in fact, the MD simulations described in this Chapter were developed only in the last 3-4 months of this thesis.

We realized a series of MD simulations on the 10bp hairpin with native sequence, using the GROMACS 5.1 computer code [18, 92]. Firstly, a structural model was generated, for the entire (hairpin + handles) molecular construct (Figure 45) to reproduce as closely as possible the molecule captured in the optical-tweezer experiments (see Fig. 17 in Chapter 3). The molecular model was built as a continuous ssDNA chain spanning from the 5′ to the 3′ ends, where the first and last groups of 29 bases were matched to two complementary 29-long ssDNA strands (*splint*), to make up the two dsDNA handles,

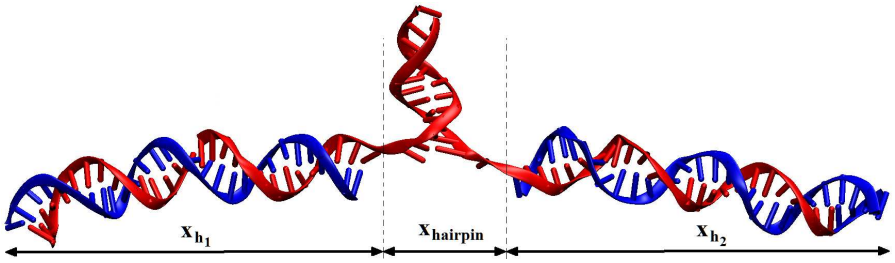


Figure 45: Molecular dynamics model of the 10 bp hairpin. The 82-nucleobase single strand making up the hairpin and half of each two handles (red), is complemented with the two splints (blue), to form the complete dsDNA handles as used in the experiments; the hairpin remains sandwiched between the two. In evidence the contribution to the total molecular elongation separately due to the handles (h_1,h_2), and to the hairpin ($x_{hairpin}$).

and the central 24 bases represented the hairpin, perfectly folded in the initial configuration.

The end-to-end distance between the C1′ atoms of the first and last bp (to be used as reference length in the foregoing) is λ_0=23 nm. The structure of hairpin plus handles, with the same native base sequence used in the experiments, was assembled in a water box of size $50\times9\times12$ nm^3 with periodic boundary conditions in the three directions, containing about 174,000 TIP3P water molecules, plus 625 Na^+ and 488 Cl^- ions, to ensure neutralization of the phosphate backbone charge, and physiological salt concentration around 0.15 M.

Equilibrium MD simulations were carried out at temperatures ranging from 300 to 360 K, and pressure of 1 atm, at constant-{NVT}. Coulomb forces were summed by shifted particle-mesh Ewald electrostatics, with real space cut-off set at 1 nm; long-range dispersion forces were also cut-off at 1 nm. We used rigid bonds for the water molecules, which allowed to keep the time step to 1 fs for both the thermal equilibration runs, and the force-pulling simulations. Typical preparatory constant-{NPT} MD runs lasted between 10 and 20 ns; force-pulling simulations were carried out for 50 ns; thermal equilibrium simulations at constant-{NVT} lasted typically 100 ns.

5.1 HAIRPIN UNFOLDING BY AN EXTERNAL FORCE

In a first type of *non-equilibrium* MD simulations, we performed simulated force-pulling experiments on this 10bp native hairpin. This was achieved by using the steered molecular dynamics (SMD) code at constant pull velocity, available in GROMACS. We fixed the center of mass of the first base-pair at one end of the dsDNA handles, and applied force by moving at constant velocity a fictitious harmonic-spring potential attached to the center of mass of the last base pair of the dsDNA, at the opposite end of the other handle. After some tests, the spring constants were set at 100 and 75 kJ mol^{-1} nm^{-2} and the pulling speed at about 20 cm/s. Such values, even if not compatible with the soft spring used in optical tweezers experiments, must be imposed to observe the hairpin unfolding over the time-scale of the MD simulation. Forces and displacements were recorded at intervals of 10 time steps, that is 10 fs, or 100-THz sampling rate.

In this way, the opening λ observed between the two opposite ends of the dsDNA handles is almost linearly increasing with time (see Figure 46 blue curves), starting from the zero-force value λ_0 =23.0 nm with the hairpin in the folded state. For a pulling velocity in the range of a few cm/s, this translates in a similarly linear opening of the hairpin, as measured by looking at the relative distance between the sugar C1′, or the backbone P atoms of the first base pair (namely, the GC pair directly linked to the two dsDNA handles, see scheme in Figure 47). As it is possible to observe in Fig. 46, the extension of the handles (green traces) remains almost constant during the simulation, except for a short initial reorientation along the pulling direction (notably, this part is neglected in the experiments because the zero force

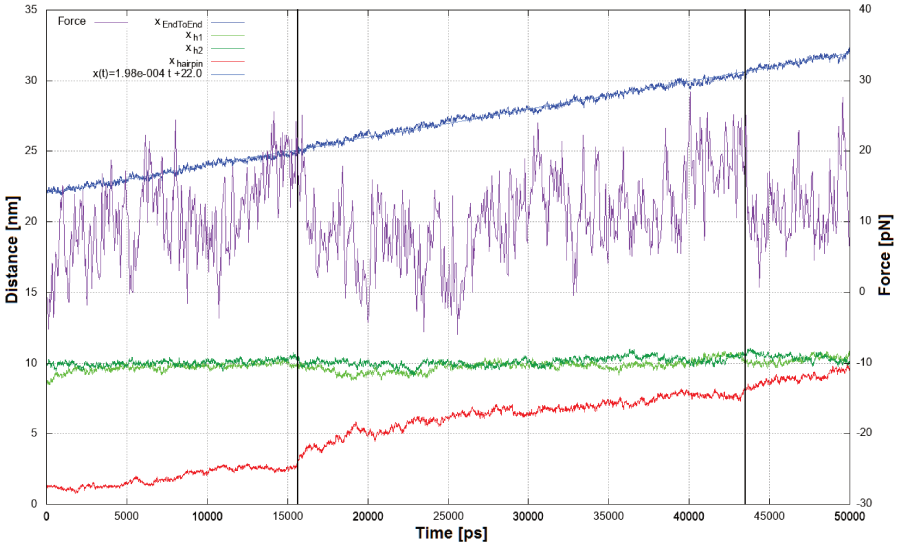


Figure 46: Time-evolution during the SMD simulation of some relevant molecular quantities: total end-to-end distance (blue) and its linear fit (light-blue); the two handle extension (green), which prove to remain almost constant at the force experienced during the simulation; the force applied at the hairpin ends (purple); the hairpin extension (red), showing a variation of the slope during the extension, in correspondence with some drop in the force (vertical black lines). These latter will be shown in Fig.47 to be a consequence of the cleavage of hydrogen bonds between base-pairs.

limit is never reached). In fact, the extension of the two dsDNA remains constant at 9.8 ± 0.4nm and 10.1 ± 0.3nm, while the hairpin (red trace) changes its end-to-end distance under the applied force.

The simulated force-pulling experiments are followed up to the complete opening of the hairpin, over trajectories of typical duration of $\sim$ 50 ns. An example of the results is shown in Figure 47, where the relative distance ("opening") between the C1′ atoms of each base pair in the hairpin is shown as a function of the simulation time (bp are numbered from 1, next to the loop, to 10, next to the dsDNA handles, see scheme in the figure inset). One first observation that clearly emerges from such a plot is that the unfolding of the hairpin does not seem to proceed by an ordered, progressive snapping of each base pair in sequence, but rather follows a kind of collective process, in which groups of bp open up simultaneously, e.g. at times $\sim$ 17 and $\sim$ 43 ns. Furthermore, the outermost bp (n. 10 in the Figure) seems to be already opened from the beginning of the simulation, and the second one follows almost immediately, after just a few ns. This observation may have interesting consequences, to be discussed in the foregoing.

5.2 TEMPERATURE-INDUCED UNFOLDING OF THE HAIRPIN

To bypass at least in part the time-scale limitations of MD, we performed a second set of simulations by picking a few "interesting" configurations from the trajectory shown in Figure 47, at times t $\simeq$7.5,

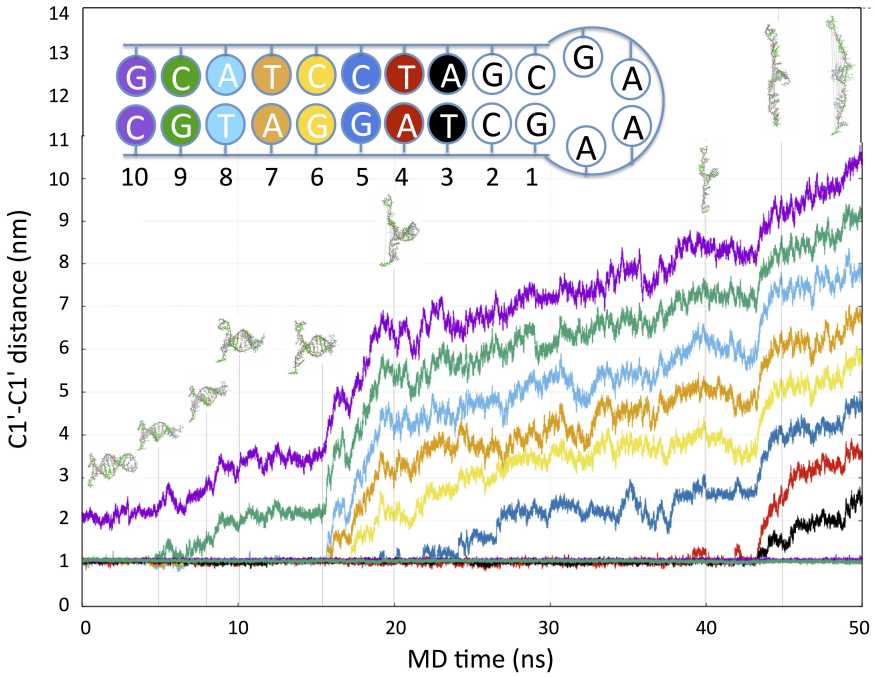


Figure 47: Results of a MD simulation at constant velocity pulling and T=300 K for the 10bp hairpin with native sequence (shown in the inset above, with base-pair numbering and color codes). The 50-ns long time traces, starting from the folded state and going to the fully unfolded, display the relative distance between the sugar C1′ atoms of each base pair, as indicated in the legend (the bp 10 being the one close to the dsDNA handles, and the 1 being the one adjacent to the loop). Above the plots, schematic snapshots give a visual indication of the average molecular configuration of the hairpin at approximately the time corresponding on the x-axis.

18, 22, 28, 35, 45 ns, corresponding to a relative opening between the opposite ends of the dsDNA handles of $\lambda \simeq$23.6, 26.1, 26.4, 27.6, 29.0, 31.1 nm. Each of these configurations were then run in a 100-ns MD simulation at constant-{NVT} with the fixed-λ external constraint (SHAKE-LINCS algorithms[65, 147]). The hairpin *quasi-static equilibrium* dynamics in such conditions may be thought of approximating the (practically) infinitely-slow pulling of the real experiment on the much faster MD timescale.

Figure 48 shows the equilibrium probability distributions of the C1′-C1′ distances for each base-pair in the hairpin, with color codes corresponding to those of Fig.47. The four panels correspond to four progressive opening values, $\Delta\lambda = \lambda - \lambda_0$ =0, 0.6, 4.4 and 8 nm. It may be noticed that at zero opening (i.e., zero average external force) the equilibrium distribution confirms the above observation, that the first base pair is constantly opened up with a C1′-C1′ distance of about 1.9 nm. At $\Delta\lambda$=0.6 nm (corresponding to about t=7.5 ns in Fig.47) the first two base pairs are spread open, and the third one is just beginning to broaden its equilibrium width. At $\Delta\lambda$=4.4 nm (corresponding to t $\simeq$28 ns) the outermost six base pairs are widely opened, while the four inner ones are still closed at their equilibrium C1′-C1′ of 1.02

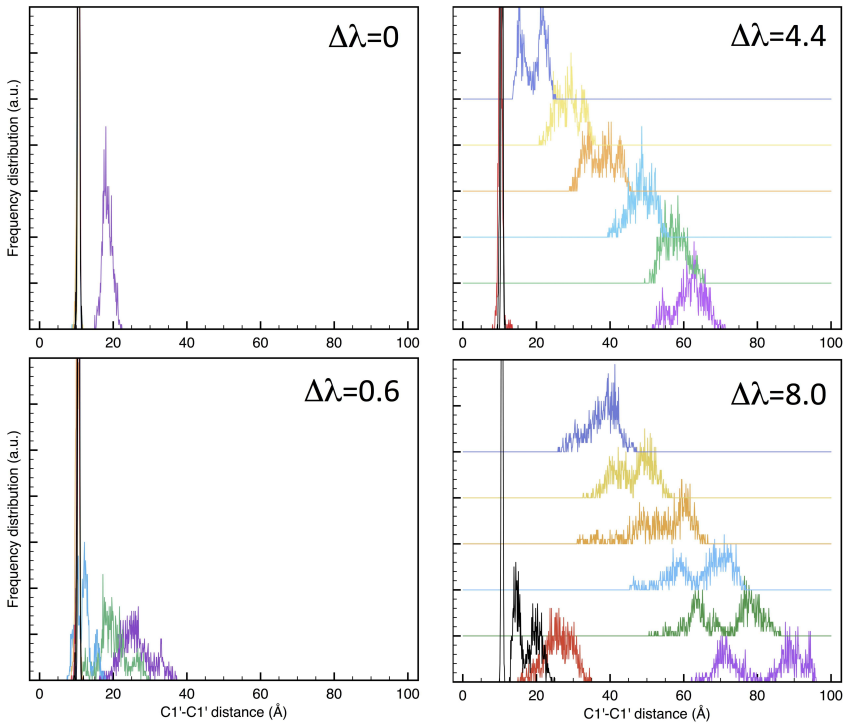


Figure 48: Equilibrium probability histograms for the C1′-C1′ distance of each base pair in the native hairpin, from 100-ns long MD {NVT} trajectories at Δλ=0, 0.6, 4.4, 8 nm. The histograms are colored according to the base-pair scheme of Fig.47; data for the two larger Δλ are shifted upwards on the y-axis for better clarity.

nm. Eventually, at Δλ=8 nm (corresponding to t ≃45 ns) two more base pairs start opening, and only the two closer to the hairpin loop are still in the closed state. Notably, at openings larger than about 3-4 nm, most base pairs display a doubly-peaked distribution, very likely indicative of the rotational flipping in-and-out of these bases about the backbone, and roughly parallel to the main hairpin axis.

Finally, in a last set of equilibrium, finite-temprature MD simulations, we wanted to test the *excitation dynamics* of the hairpin. These simulations were run for 100 ns at constant-{NVT}, for all the values of fixed Δλ above, by increasing the temperature of the simulation in steps of 10 K above room temperature. In Figure 49 we present a subset of the results, namely the data for one particular value of the opening, Δλ=4.4 nm, and for three temperatures T=300, 320, 340 K (other data have a closely similar behavior). At such opening, each 20 K increase in temperature corresponds to an applied force Δf ≃18.5 pN, that is a value comparable to the experimental coexistence force f_c. Therefore, by running MD at 320 and 340 K under fixed opening, we are simulating the effect of injecting once, or twice an amount of energy Δf · Δλ: for a fixed Δλ, this would in some way be equivalent to proportionally increasing the effective stiffness of the optical trap.

The Figure shows the probability distributions for the base-pairs 10 to 4, the innermost 3 pairs remaining closed at any temperature; the traces in color correspond to T=300 K (blue), 320 K (black) and 340 K (red). It is noted that at room temperature, the energy injected to

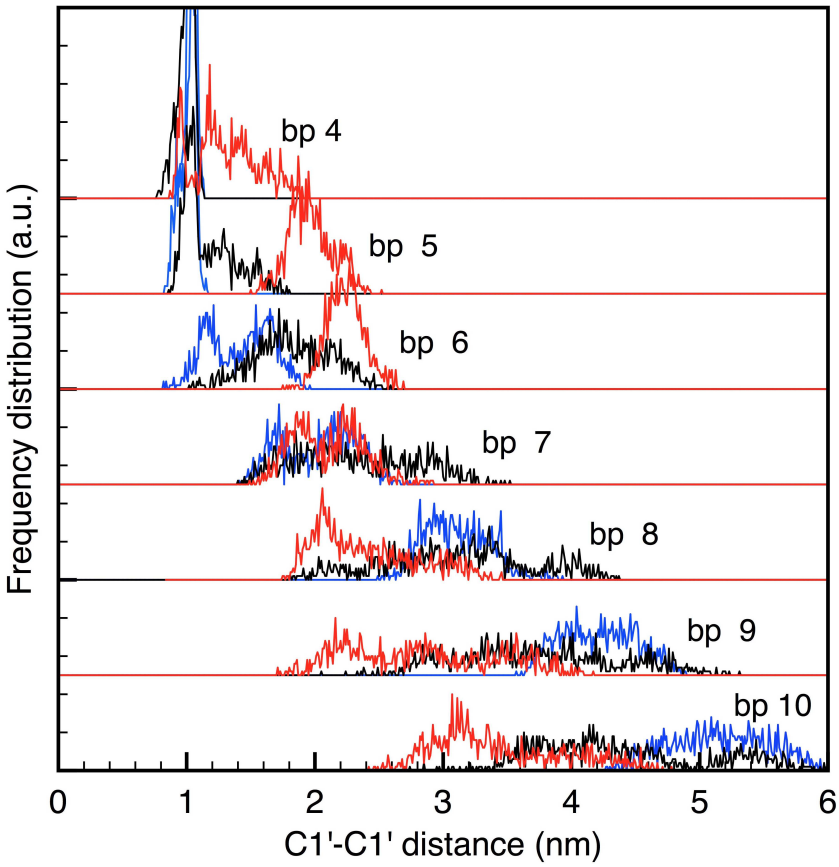


Figure 49: Equilibrium probability histograms for the C1′-C1′ distance of each base pair in the native hairpin (numbered according to the scheme in Fig. 47), from 100-ns long MD {NVT} trajectories at fixed $\Delta\lambda$=4.4 nm, and different temperatures: T=300 K (blue), 320 K (black) and 340 K (red).

maintain the hairpin at that opening of $\Delta\lambda$=4.4 nm is quite unevenly distributed among the outermost base pairs (6 to 10, with the 5 being only slightly excited). These display average amplitudes decreasing towards the loop end; the width of the distribution is nearly doubled for the outermost base-pair 10, compared to the other ones. As the injected energy (i.e., the simulation temperature) is increased, the distributions tend to become more even (the centroid of the distribution decreasing for bp 8-10, and increasing for bp 4-7), while at the same time each distribution covers a very broad range of base-pair opening, now spread over about 2 nm width for the outermost pair. We interpret these results as being the effect of an increased cooperativity in the unfolding transition, at high values of force: while at low forces the base-pairs tend to open individually, or in small batches, at high applied forces the unfolding rather appears to occur by a simultaneous opening of the entire hairpin.

5.3 FINAL SUMMARY

In conclusion, the MD simulations on the hairpin+handles molecular constructs confirm that the dsDNA handles offer a very small contri-

bution to the total molecule extension during force-induced stretching, and that the hairpin displays a variable degree of cooperativity in the folding/unfolding transition.

One important information for the analysis of experimental data, was the observation that even at nearly zero force the first base-pair (the one to which the dsDNA handles are covalently attached) is always already opened. This observation, coherent with other experimental observations by other groups, actually reduces the size of the already very short hairpin, allowing to bring coherence among the experimental data obtained with the hopping and the force-pulling methods.

It was shown that during the structural transformation, at lower forces the hairpin appears to unfold in a sequential way, but with groups of bases opening up together, while at higher forces the hairpin tends to open up in one collective snapping of the bonds between base pairs. For a given displacement λ, a larger force translates into a stiffer optical trap, and such an increase in cooperativity with stiffer coupling is in agreement with the theoretical predictions of previous works in our group [101]. The transition from an additive to a collective unfolding may also be interpreted in terms of an increasing "friction effect" that builds up between the closed base-pairs, which must overcome a twist elastic barrier, at the same time as the chemical bond-breaking barrier [66]; however in the present case this effect would be driven by a variable force, rather than by a variation in the polymer physical length.

In the next Chapters we will move from this simple DNA-hairpin system, to systems that describe full double-stranded DNA fragments in the presence of single- and double-strand breaks. This new set of studies will provide microscopic information not easily accessible to experiments, will allow to estimate limits to the life-time of the defects, and possibly suggest ways to experimentally perform some tests on such defects.

6 LINKER-DNA SIMULATIONS

In this Chapter we describe the effects of backbone breaks (SSBs and DSBs) in a fragment of dsDNA, studied by MD simulations. A better comprehension of the mechanics of SSBs and DSBs should help in understanding the sequence of events leading to the fracture of DNA, as observed in chromosomal translocation. Moreover, it may shed light on some features in the functions of the defect recognition mechanism by the repair proteins ("what exactly these proteins are looking for?"). Finally, it can be used to improve theoretical models of the rupture statistics under irradiation, such as the ones developed in our lab [102, 103, 117] to describe random break formation in DNA bundles. We start by constructing a MD structural model for the DNA system, and verify its stability; next, we introduce phosphate bond cuts in the backbone structure, to simulate the presence of the strand breaks as could be induced by ionizing radiation; finally, we compare the mechanical properties of the different models, to extract information about infrared spectra, principal bending/torsional movements, and full details of the molecule fracture dynamics.

6.1 MOLECULAR STRUCTURES OF DAMAGED LINKER-DNA

In the initial phase of the thesis work we used the make-NA server (`http://structure.usc.edu/make-na/server.html`) to generate DNA structures; in the follow up of our studies, we built a small utility code that can generate arbitrary dsDNA configurations of any given sequence, also including curved, bending and torsion pre-stressed configurations. We built two reference double-stranded DNA fragments of 31 base-pairs in the B-form: one with a random sequence, and the other a TATA-sequence. The first is taken to be representative of the exposed "linker" DNA, between two nucleosomes in the chromatin 10-nm fiber; the other was used as a reference in just a few simulations, to check against some possibly sequence-dependent properties. In the initial simulations, DNA fragments were solvated in a water box of size about 4x4 nm^2 in cross section, and ~12 nm in length, with periodic boundary conditions in the three directions, containing about 6,000 TIP3P water molecules, and enough Na^+ and Cl^- ions to ensure neutralization of the phosphate backbone and a physiological salt concentration around 0.15 M (Figure 50).

The preparatory annealing runs are composed by an initial energy minimization by conjugate gradient of the raw structure; then, a first MD run at constant-{NVT} at 310K is performed by fixing the water positions, just to reduce internal stress in the DNA molecule. Then, a 10 ns run at constant-{NPT} (target pressure 1 atm) is done with fixed DNA, to allow relaxation of the water box and obtain the correct density. On the output configuration, we verified the water density, ion

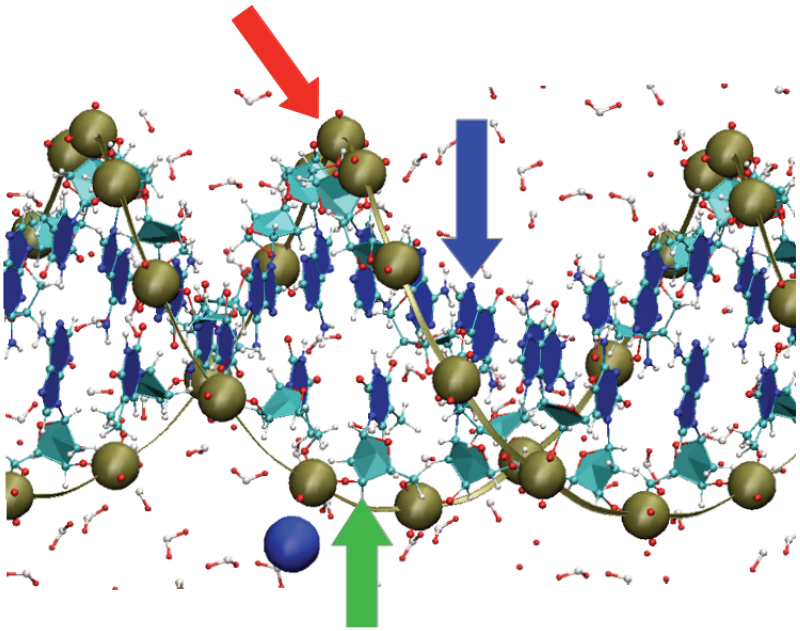


Figure 50: Initial MD configuration in evidence the DNA phosphate group in the back-bone (red arrow), the pentose sugar (green arrow), the bases (blue arrow) and the 3-site molecule of water (withe and red dots) around the molecule.

distributions, and the general DNA structure comply with the expectations from the experiments and previous literature. As described in Chapter 4, in all MD simulations we used the CHARMM-27 force field with its extensions for nucleic acids.

The final conformation of a DNA sample is the result of the three-dimensional organization of its base-pairs. We performed a longer {NPT} run to check the behavior of the structural parameters. By taking as reference the phosphate on the two backbones, and picking three atoms for each base (to identify the plane in which the base lies) we have been able to measure important structural information, such as: the inter-phosphate distance, base-pair distance, twist angle between successive bases, and verify the agreement between the model and the experimental observations, as shown in Table 6 and Figure51; in the plot are reported only the results relative to the random DNA sequence, that is the one used for further developments of the research. The values for the observed quantities are in agreement with the B-DNA conformation. Especially we found, for both models, $\sim$ 34 deg as average value for the twist, corresponding to 10.6 bases per turn around the DNA helical axis, in perfect agreement with the experimental measurements; also the P-P distance and the inter-base parameters offer good comparison with the experiments.

Once having verified the reliability of the configuration and stability of the force field, we introduced in the random sequence some different cuts in the sugar-phosphate backbone, between the bases 16 and 19 on one strand, and between 13 and 16 on the opposite strand. In this way, we obtained one SSB configuration, and three DSB configurations, in which the two opposing SSBs are respectively spaced

Table 6: B-DNA structure parameters.

	bp dist [Å]	P-P dist [Å]	Twist [°]
MD model	3.5 ± 0.2	6.8 ± 0.3A	34 ± 14
TATA model	3.6 ± 0.3	6.7 ± 0.3A	34 ± 21
Experiments	3.32	7.0	34.3

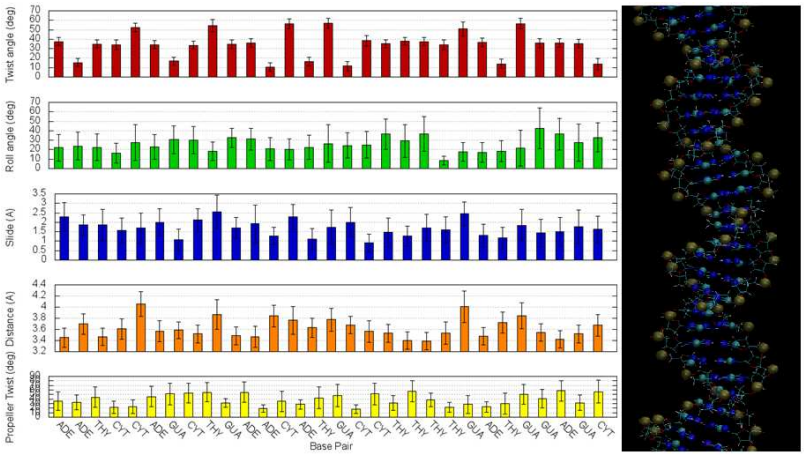


Figure 51: Time average, with relative error bar, for different structural parameters of DNA simulated with the CHARMM-27 force-field parameters. On the horizontal axis are listed the base-pairs that compose the sequence, on the y-axis from the top are presented: the twist angle, roll angle, slide displacement, base pair distance and the propelle- twist angle. On the right, image obtained with VMD of the DNA fragment; in evidence, the atoms used to identify the above parameters.

by 1, 2 or 4 base pairs (DSB 1-bp, 2-bp, 4-bp; see Figure 52). The base-pairs comprised between the two cuts of the three different DSB configurations are:

- *DSB 1-bp* A_{16}–T_{16}
- *DSB 2-bp* $T_{17}A_{16}$–$A_{15}T_{16}$
- *DSB 4-bp* $T_{19}G_{18}T_{17}A_{16}$–$A_{13}C_{14}A_{15}T_{16}$

We have chosen the random sequence rather than the poly-TA sequence because area with a mixture of G-C and A-T are clearly more representative of the active gene area. In fact, A-T base-pairs form only two hydrogen bond, compared with the three in G-C, so poly-TA regions are easier to denaturate;also for this reason they are usually found at binding site of a transcription factor. No further attention was given to the sequence dependence in the definition of the strand breaks, although this can be expected to have some (quantitative, rather than qualitative) influence on the results. For the strand-break terminations we used the a standard 5′-OH and a 3′-phosphate (3PHO) from the CHARMM library. We also prepared different terminations, a 5′-phosphate (5PHO) and a 3′-phosphoglycolate (3PPG); or a 5′-aldehyde (5ALD) and a 3′-phosphate (3PHO). These are (among

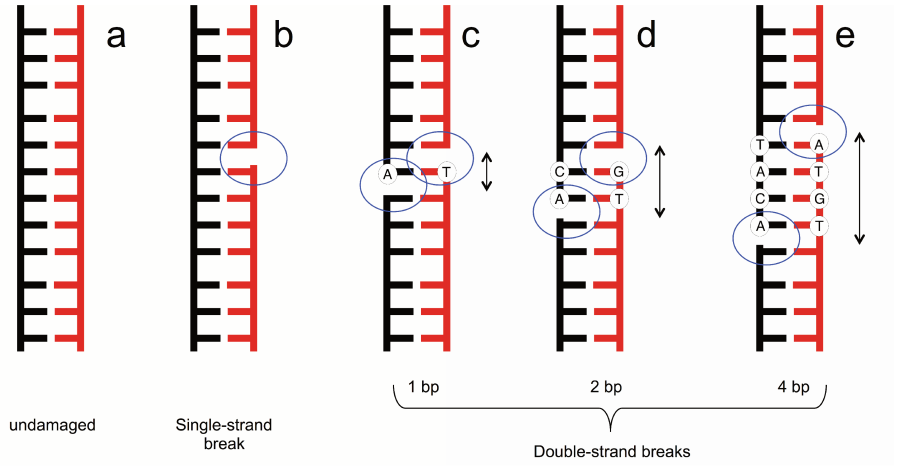


Figure 52: Schematic of the SSB **(b)** and 3 DSB defects **(c-e)**, compared to the pristine DNA **(a)**. Snapshots **(f-h)** of the molecular configurations of the DSBs, with the positions of the opposed strand breaks indicated by arrows. DNA backbone represented by a ribbon, and bases by polygons.

many others) common outcomes of backbone cleavage, in particular, the second one forming in presence of O_2 and the third one being a typical product of $OH^{\bullet}$ attack [142]. However, these were not actually used in the foregoing and may be used in a further extension of the studies presented here.

As in the preparatory simulations, DNA fragments were solvated in water box with periodic boundary conditions in the three directions. However, in this case depending on the type of analysis we were interested, the box cross section ranged from about 4x4 to 6x6 nm^2, and the length from 12 to 22 nm. Using the explicit solvent this has mean adding about 6,000 to 48,000 TIP3P water molecules plus Na^+ and Cl^- ions to ensure neutralization of the phosphate backbone and a salt concentration around 0.15 M (Figure 53), for this reason when as in the case of the vibrational spectra analysis we have preferred to use smaller box.

Again, as in the undamaged fragment, the above protocol to ensure an initial relaxed configuration at 310 K and 1 atm has been repeated. We used in most cases rigid bonds for the water molecules, which allowed to push the time step to 2 fs for the thermal equilibration runs, and to 1 fs for the force-pulling simulations. Typical preparatory MD runs lasted between 10 and 20 ns, while force-pulling and thermal stability simulations extended to 100 ns (or some time after the breaking point, if breaking occurred earlier). Standard Ewald-sum electrostatics was used for the Coulomb forces, with a real-space cut off radius of 0.7 nm. Long-range non-bonding forces were summed up to a cut off radius of 1 nm. Vibrational spectra were obtained from the Fourier transform of the velocity autocorrelation function, during constant-{NVT} MD runs of 10 ns with a time-frame sampling of the trajectory every 10 steps. Steered-molecular dynamics (SMD) simulations instead require a substantially longer water box, to avoid self-interaction of the polymer once in the over-stretched configura-

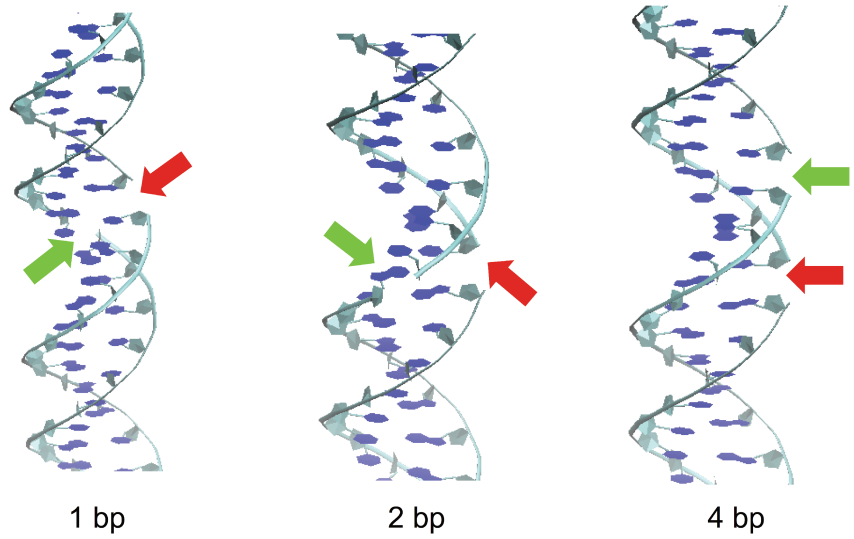


Figure 53: Molecular representation of the three double-strand break configurations after molecular dynamics equilibration at 310K. DNA backbone is indicated by continuous ribbons, DNA bases are represented by cyan pentagons (ribose) and blue heterocycles (generic bases). Red and green arrows indicate the position of the strand breaks.

tion; also, longer simulation times are required in order to keep as low as possible the pulling speed.

6.2 VIBRATIONAL SPECTRA

One main objective of this study was to identify mechanical signatures of the DNA response to the different kind, and arrangement of strand breaks. The velocity autocorrelation function provides information about the fast atomic vibrations of intramolecular bonds, which can be experimentally probed by coupling infrared and Raman spectroscopy. By comparing the vibrational spectra among the various undamaged and damaged configurations, it was hoped that some clear signature of the different types of defects could be identified. However, as shown by the spectra collected in Figure 54, such results are far from conclusive. By comparing the spectra for the different configurations with experimental peak assignments from the spectroscopic literature [15], it can be seen that strand breaks (even at the equivalent high linear density, 1 break every ~10 nm, used in the simulations) have but a minor effect on the peak frequencies, as well as on their relative intensities.

As it could be expected, the strongest effect is seen on the phosphate groups, and more clearly as a split peak for the DSB 1-bp, since in this case the two cleaved P-groups are closest to each other. However, already for the 2-bp the effect is reduced to a mere increase in the peak intensity without frequency shift, and for the 4-bp the effect is unnoticed. Moreover, such changes seem to affect only the symmetric stretch frequency around 1010 cm^{-1}, but not the antisymmetric stretch at 1240 nor the O-P-O bending at 865 cm^{-1}. The other features

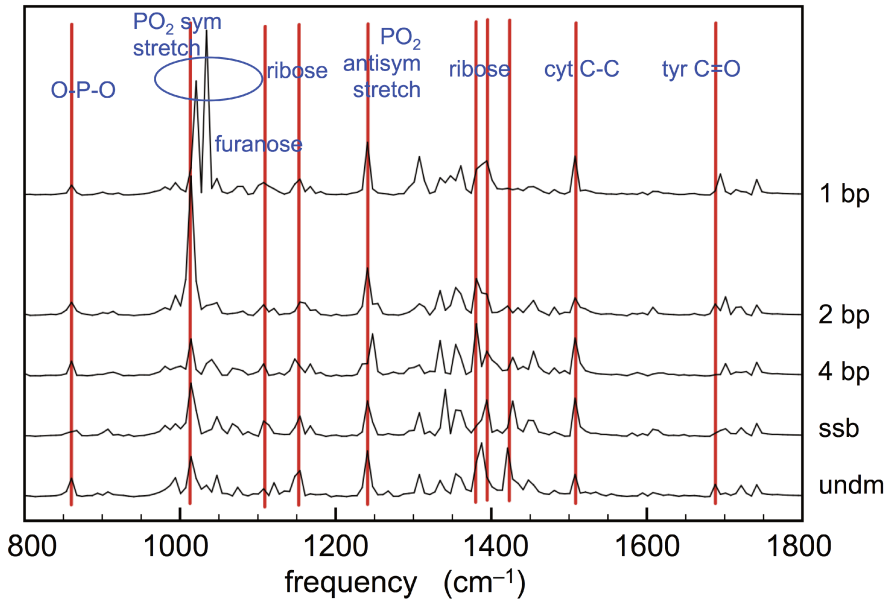


Figure 54: Vibrational spectra in cm^{-1} units, from the velocity autocorrelation function at 310K, for the undamaged, SSB, and DSB at 1, 2, 4 bp distance. Vertical red lines indicate experimental peak assignments, labeled in the row above.

which seem to be moderately affected by the presence of the strand breaks are those relative to the ribose. More evidently, the broad signal at 1390 cm^{-1} is split in two sub-peaks and its relative intensity is slightly lowered; the other peak near 1150 cm^{-1} shows some minor variation. Similarly minor variations are observed for the C-C stretching of cytosine and the C=O stretching of thymine. However, these results altogether do not seem enough to warrant discrimination among the various defects, given the much lower concentration of SSB and DSB found even in heavily irradiated DNA.

6.3 ESSENTIAL DYNAMICS

Essential dynamics was extracted from the same MD trajectories used for the vibrational spectra, with the help of the plug-in included in the GROMACS code. This analysis allows to describe the slower dynamics of the DNA fragment, complementary to the fast oscillatory degrees of freedom already observed by the vibrational spectra. Once projected on the eigenvector spectrum, the covariance matrix concentrates almost 90% of the total weight in the first few eigenvectors, while the others, as expected from previous studies, describe only harmonic oscillation of constrained degrees of freedom. Figure 55 actually confirms that, for any DNA configuration in the present study, the analysis can be practically restricted to the first two eigenvalues, with a minor contribution from the eigenvalues 3 to 5, and negligible contribution from the remaining ones.

Starting from the eigenvectors $\mathbf{v}_i$ (three columns of the matrix $\mathbf{V}$ for the Cartesian components of each atom), principal coordinates (PC) can be introduced as $p_i(t) = \mathbf{v}_i \cdot (\mathbf{r}_i - \langle \mathbf{r} \rangle)$ [63], with the purpose

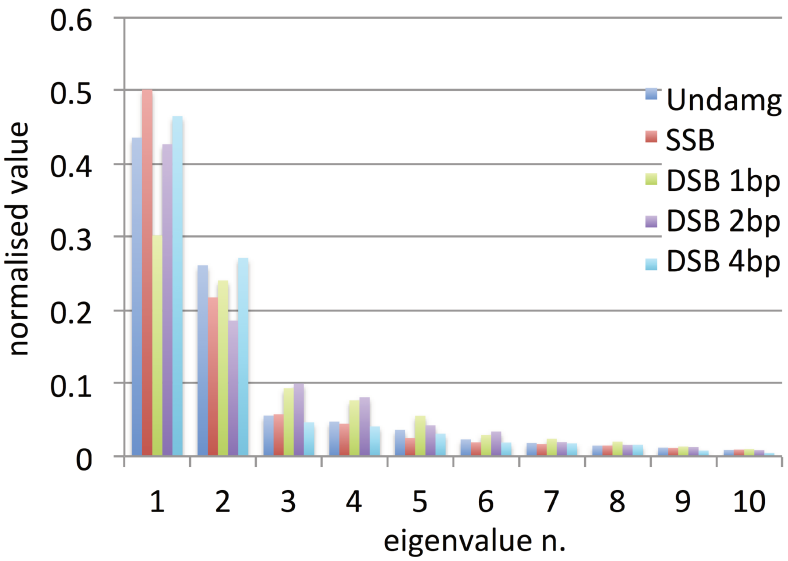


Figure 55: Histogram of the weight of the first ten eigenvectors of the covariance matrix, for the undamaged DNA and the four different strand-break defects, obtained from a MD run of 5 ns at T=310 K. The sum of the weights is normalised to 1 for each configuration.

of highlighting particular components of the large-scale movements. Defined the main axis of the DNA as the z direction, we calculated the following collective coordinates in order to describe and quantify the deformation produced by each eigenvector, respectively, as *bending* and *torsion*:

- $q_i(t) = \sqrt{\left(v_i^x \Delta r_i^x\right)^2 + \left(v_i^y \Delta r_i^y\right)^2}$ represents the global displacement projected in the plane perpendicular to the z-axis and could be used to evaluate the bending.
- $t_i(t) = -v_i^x \Delta r_i^y + v_i^y \Delta r_i^x$ with $\Delta r_i^\alpha = r_i^\alpha(t) - \langle r^\alpha \rangle$, $\alpha = x, y$ the time-dependent deviation from the mean of the Cartesian components of $\mathbf{r}(t)$ to detect torsional movements could be used to evaluate the torsion about the DNA axis

In Figure 56(a,b) we plot the time-averaged values of q_i (left panels) and t_i (right panels) for each atom i, arranged according to the respective z-coordinates. The results shown in panels (a,b), respectively, are relative to the two PCs calculated for the first and second eigenvalue of the covariance matrix; a reference DNA structure is also shown below each figure, to facilitate the interpretation. Quite generally, the first two eigenvalues appear to contain a very similar information for any defective DNA configuration.

The projected displacement q_i appears as a global bending about the central region of the molecule, where the strand breaks are accumulated (note that since the q_i are defined as squared quantities, the rather sharp 'V' observed at about $1/4$ and $3/4$ of the length along z is actually an indication that the displacement crosses over to the negative axis; in other words, what is represented is actually the modulus of $|q_i|$). However, for the DSB 1-bp configuration the bending at the center is very straight, resembling a rigid flexion about a

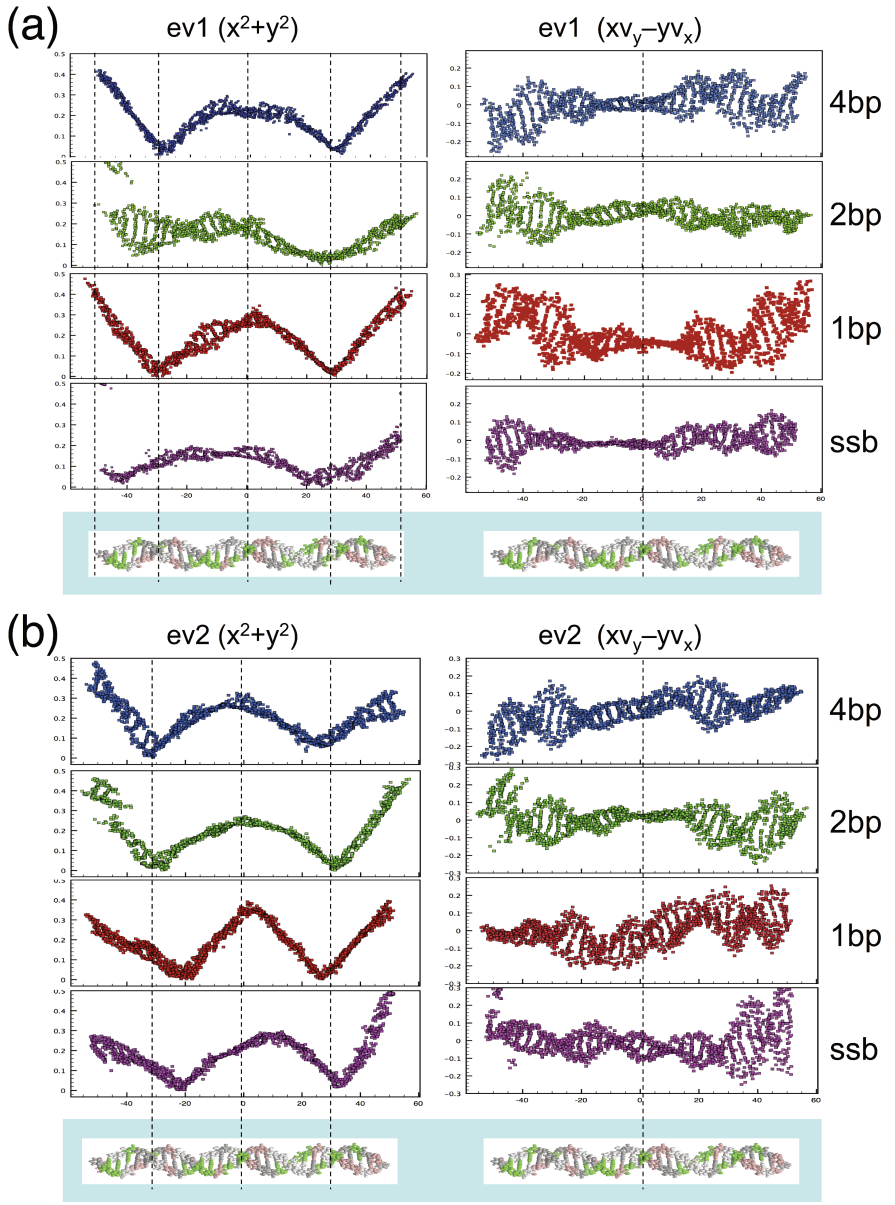


Figure 56: Atom-by-atom representation of the two collective coordinates for the damaged-DNA configurations. The values of q_i (left, indicated as "x^2+y^2") and of t_i (right, "$xv_y - yv_x$") for each atom are arranged along the respective z-axis coordinate. **(a)** First eigenvalue of the covariance matrix. **(b)** Second eigenvalue. Below each panel, the reference DNA configuration is sketched within the grey panel; dashed lines are guides to the eye.

hinge, whereas for the other configurations it appears more smooth and continuous about the center. The behaviour of axial torsion t_i is also rather similar for the four defective structures. For the first eigenvalue, the quantity t_i shows a sort of 'necking' in the central region (more pronounced for the SSB and the DSB 1-bp), while the torsion about the z-axis looks more continuously distributed along the whole z-length, for the second eigenvalue.

Anyway, also this type of long-wavelength oscillation analysis based on ED seems to be of little help in providing characteristic signatures of the different defects, as they could be appreciated by some ex-

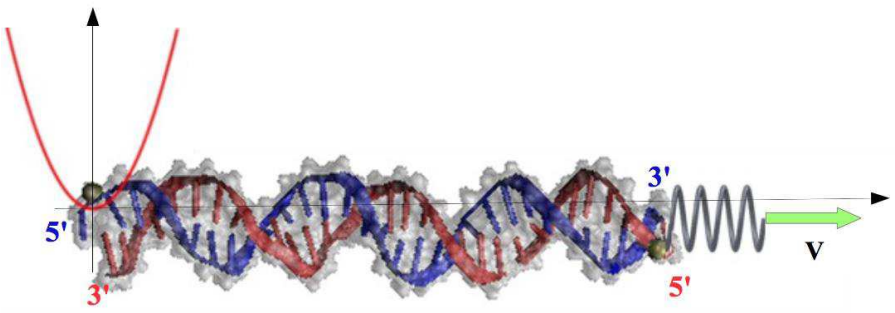


Figure 57: Schema of the set-up in steered molecular dynamics simulation: the DNA fragment is bound by its 5′ termination to a fixed point by an harmonic potential and to a "dummy atom" moving at constant velocity by a spring.

perimental technique, with possibly the only exception of the sharp "hinge-bending" of the DSB 1-bp, which appears distinctly different from the other collective movements.

6.4 SIMULATED FORCE SPECTROSCOPY

Steered molecular dynamics was performed on the five fragment configurations (undamaged, SSB, and 3 different DSBs) with the constant velocity plug-in available in both NAMD and GROMACS. Both ends are linked by a virtual spring to the 5′ end of each DNA fragment: one end was fixed to a rigid wall; while the opposite was attached to a moving spring via the last P atom (Figure 57). After some tests, the spring constant was set at 1,000 kJ mol^{-1} nm^{-2}, or 1,660 pN/nm. Pulling velocities in the range 12 to 1 cm/s were used, with most SMD simulations being carried out at the intermediate speed of 4 cm/s. Forces and displacements were recorded at intervals of 5-10 time steps.

The DNA elongation (displacement) is measured at each time step by computing the distance between the positions of the terminal P atoms on each of the two 5′ ends of the fragment, therefore it is strictly linear with time while displaying a very narrow fluctuation. On the other hand, during the constant-velocity SMD simulation and especially at extremely low pulling speeds, the force fluctuates very much. Therefore, a special averaging procedure was devised to minimize the noise, in three stages:

- Stage 1: (a) the output force is recorded at 5 fs intervals, while the displacement (difference between the absolute position in space of the two P atoms at which the virtual springs are attached) is recorded at 20 ps intervals; (b) each position is obtained as the average of two successive displacements (40 ps), over which 8,000 force values are also averaged. For a 100-ns simulation, 2,500 force-position pairs (fpp) are thus obtained.

- Stage 2: to further reduce the noise of the force-position curves, moving averages are calculated with windows of different width, namely 0.2, 1, 2, 10, 20 ns; a width of 0.2 ns averages 5 fpp, a

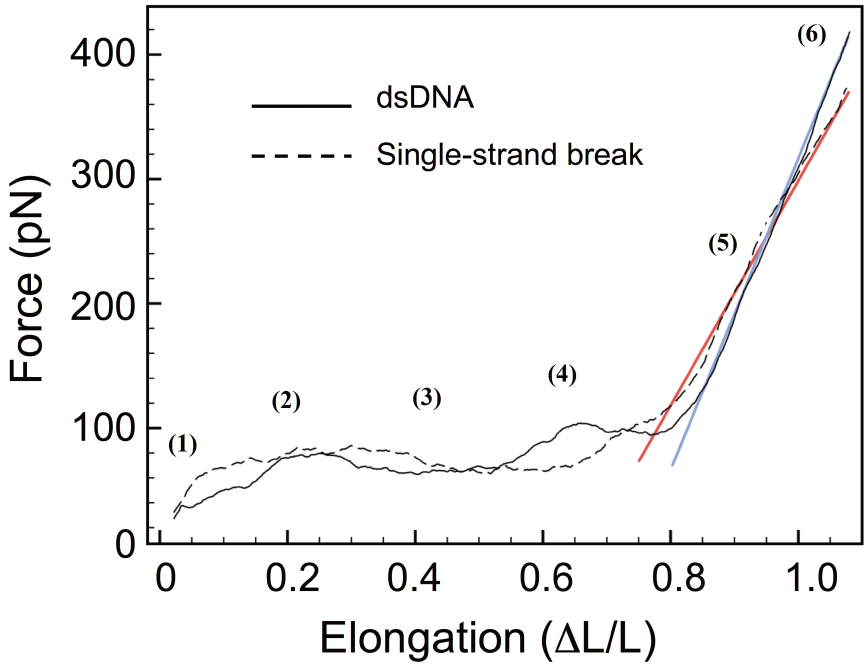


Figure 58: Force-elongation plot from a MD simulation at T=310 K, for undamaged (continuous curve) and SSB (dashed) DNA configurations, pulled along the main *z*-axis; spring constant 1,660 pN/nm, pulling velocity 4 cm/s. The straight lines represent the fit to extract the respective elastic modulus (blue=undamaged, red=SSB). Position 1-6 are marked to compare the curve to the molecule configuration in the following figure.

width of 1 ns averages 25 fpp, and so on. The 2-ns window appears to offer the best compromise between reducing the noise while preserving the essential features of the force-displacement curve.

- Stage 3: the 2-ns averaged curves from at least three SMD simulations with different initial conditions are further averaged, at equal values of displacement, for each DNA configuration.

In Figure 58, we report the averaged force-displacement curves for the undamaged 31-bp DNA (full curve) and the SSB-damaged DNA (dashed). After the initial rise (here different from the standard worm-like chain shape, since we start from a constrained, nearly straight fragment), the force oscillates for both systems between 60 and 100 pN, over a large plateau of displacement up to about 70% relative elongation. As shown in Figure 59 this plateau in the undamaged sequence corresponds to the transition from the B-DNA to an overstretched conformation. Experimentally, such a plateau is observed at a constant force of about 65 pN, not dissimilar from the average values of our simulations. After about 80% relative elongation, the DNA fragment is in both cases completely extended and untwisted, therefore the strictly entalphic part of the stretching appears as a nearly linear force-displacement relationship, whose slope (blue and red lines) is related to the apparent Young's modulus E of the molecule. By taking a DNA diameter of $\simeq$ 2 nm, we obtain estimates

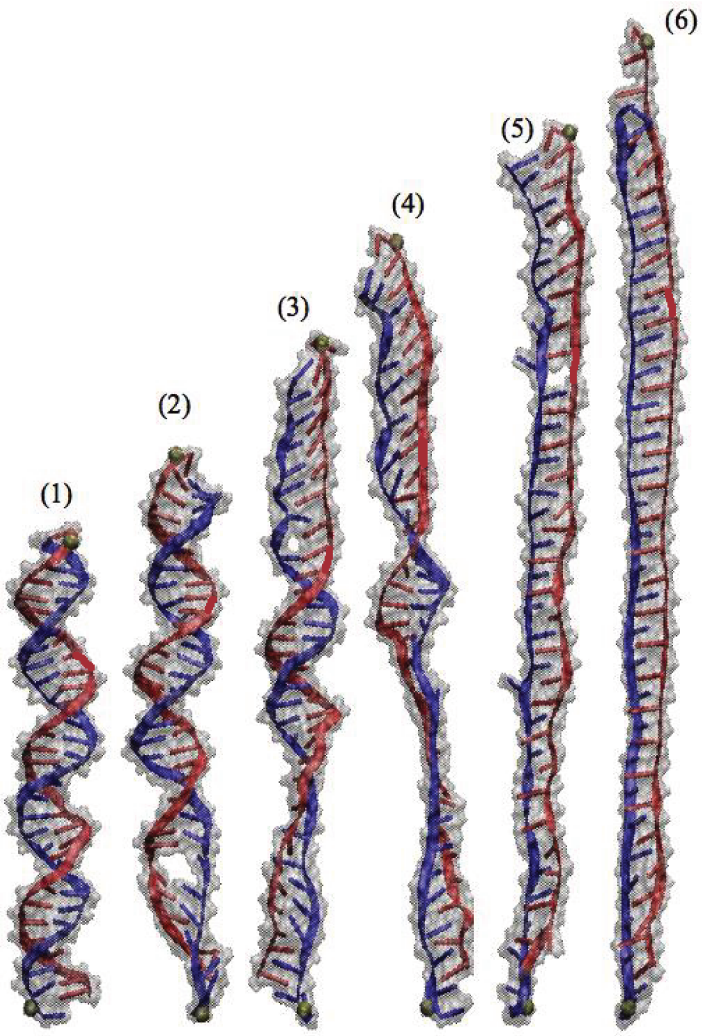


Figure 59: Configuration of the undamaged DNA fragment at different elongation (1-6). The comparison with the relative positions in the force-elongation cure show how the plateau correspond to a transition from B-DNA to an overstretch conformation.

of E=330 MPa and E=270 MPa for the undamaged and the SSB-DNA, respectively. It may be worth noting that the apparent stiffness of the SSB-DNA is quite larger compared to about half that of the pristine DNA, as one could have simply guessed by considering that the former is held by only one intact strand. It is likely that the combined effect of torsion and bending forces about the SSB defect helps in this case, in relieving part of the tensile stress, thereby leading to a higher apparent modulus.

In Figure 60, similarly averaged force-displacement curves are reported for the three different DSB-containing DNA fragments. However, because of the too large difference between separate simulations with different initial conditions, step 3 of the averaging was not carried out in this case; therefore each curve represents one individual simulation. The three plots show a similar behavior, with a complex force spectrum characterized by subsequent peaks, followed by a sharp drop to zero force, signalling the ultimate break-up of the 31-bp DNA into two separate fragments. The 1-bp, 2-bp and 4-bp

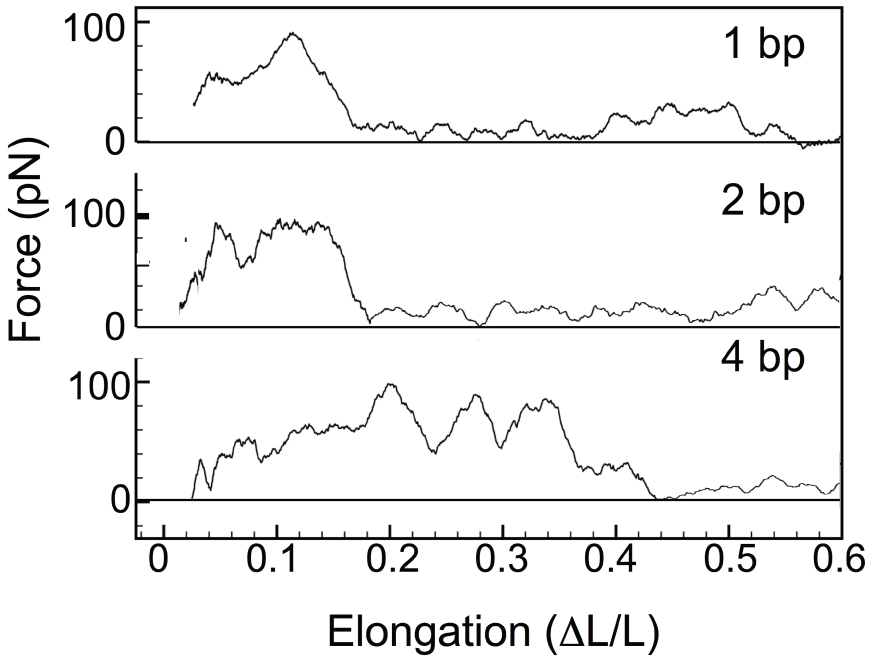


Figure 60: Force-elongation plots from MD simulations at T=310 K; pulling direction along the main z-axis; spring constant 1,660 pN/nm, pulling velocity 4 cm/s. **(a)** Undamaged (continuous curve) and SSB (dashed) DNA configurations. The straight lines represent the fit to extract the respective elastic modulus (blue=undamaged, red=SSB). **(b)** DNA with different DSB: from top to bottom, 1-bp, 2-bp and 4-bp break distance.

cases display respectively one, two and three main peaks, of variable height, reaching at different times (or elongations) a maximum value of $\simeq$100 pN. To a first guess, these values could be broadly identified with the average force necessary to break one individual base-pair. This is certainly reasonable for the 1-bp configuration, for which a graphical rough estimate of the integral under the force-displacement curve gives a mechanical work of about 12 kcal/mol, in broad agreement with the results of quantum chemical calculations for breaking one A:T pair [153]. However, for the other two cases in which the DSB occurs across a larger base-pair separation of 2- and 4-bp, the mechanical response of the damaged DNA under tensile force is already quite more complex.

In the 4-bp simulations, smaller peaks are also observed, both before and after the four main ones (the last one being considerably smaller). Notably, for both cases we observe yielding to occur by a sort of stick/slip mechanism, i.e. a relative sliding of the base-pairs along the z-axis (the direction of the pulling force) with H-bonds being broken and reformed also in "violation" of the standard Watson-Crick pairing rules. Therefore, the broad peaks observed in the force-displacements plots appear to indicate *collective* rather than discrete events, such as individual bond breaking.

The first part of the force-displacement plot in Figure 60, from 0 to about 16% elongation, corresponds to the progressive breaking and reforming of H-bonds within the four base-pairs comprised in

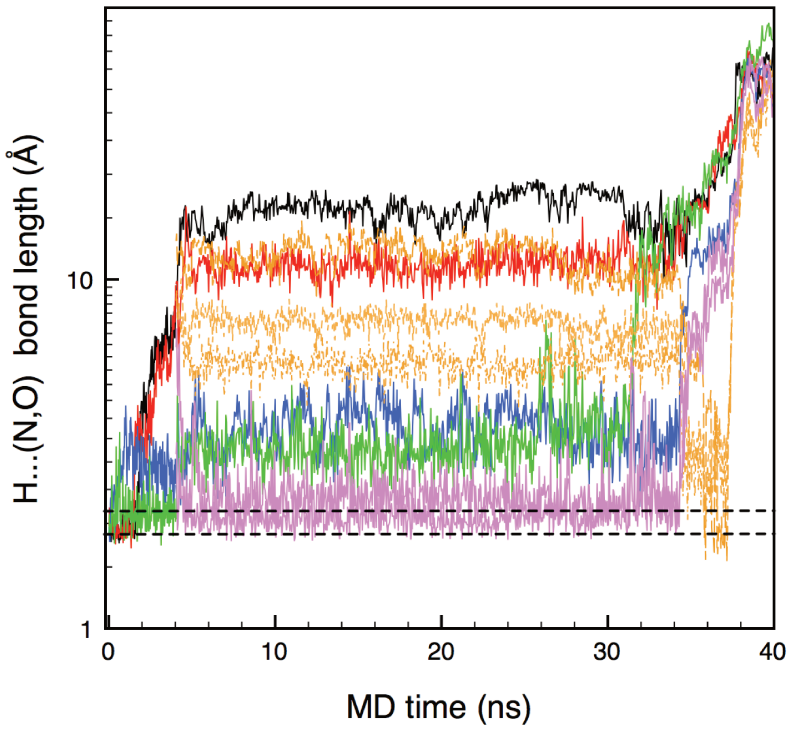


Figure 61: Plot of selected hydrogen-bond lengths between the 4 base-pairs comprised between the backbone cuts of the DSB-4p configuration, during a MD simulation at T=310 K, with spring constant 1,660 pN/nm, pulling velocity 4 cm/s. The two horizontal dashed lines delimit the H-bond length comprised between about 1.8 and 2.1 Å.

the DSB. Figure 61 shows plots of a few selected H-bond lengths between the four base-pairs, during one MD simulation at T=310 K. For reference, H-bond lengths for the canonical Watson-Crick base pairs are typically in the 1.7-2.1 Å range at zero temperature [153]. It can be seen that different groups of H-bonds (different colors) come into service as long as the tensile force pulls the DNA. The first H-bond to break, around t=0.4 ns, is the G_{18}-C_{14} H...O, followed right after (t ~1-2 ns), by both the T_{19}-A_{13} H-bonds (see also the small peak of 40 pN at ~4% elongation in Fig. 60). At the other end of the DSB, the bonds in A_{16}-T_{16} hold until t=4.5 ns, when they split up. In the meantime, new H...O and H...N bonds are formed, between the T_{19}-C_{14} and G_{18}-A_{15}, non-Watson-Crick base pairs.

Figure 62 shows details of the switching of the T_{19}-H3 hydrogen, which changes bond from the A_{13}-N to the C_{14}-N, around time t=4 ns. Such bonds span a long DNA stretching time, until about t=34 ns (~36% elongation). Figure 63 shows, for the two bonds between H1...N3 and H21...O4′ of G_{18} and A_{15}, the geometric parameters: H-bond length; donor-acceptor distance; and H-d-a angle.

Such a peculiar yielding by progressive sliding can be very clearly appreciated in the case of the DSB 4-bp. During this large time span (~4 to 34 ns), a sequence of collective events takes place, while the two strands continue to slide with respect to each other. This may suggest some conceptual similarity with the mechanism of a dislocation glid-

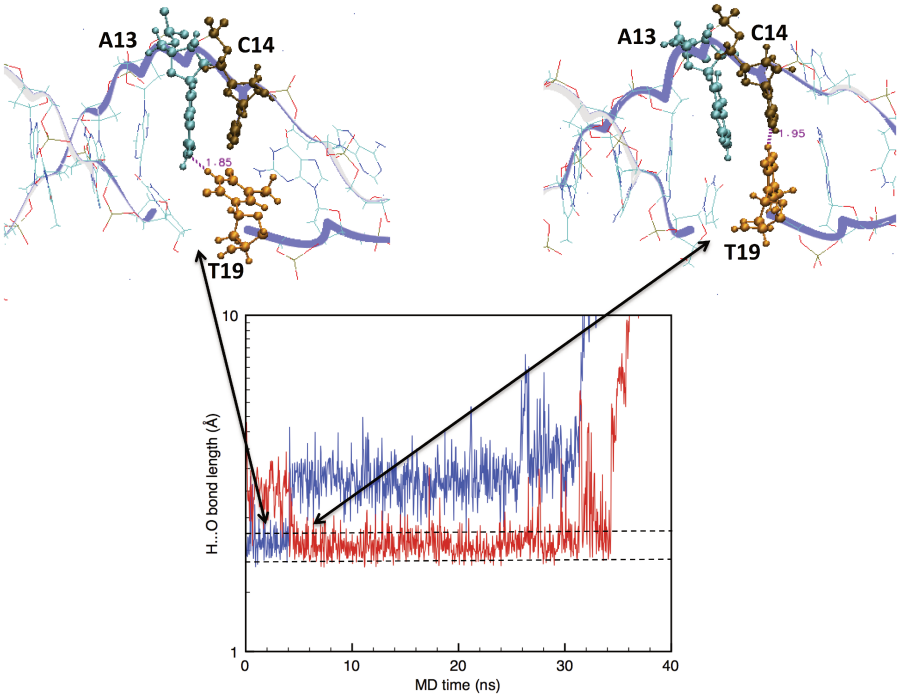


Figure 62: Time evolution of the H-bonds formed by the T_{19}-NH, initially with the N7-A_{13} (blue trace), and then with the N3-C_{14} (red trace), thus forming a non-Watson-Crick pair about one of the strand breaks, during a force-pulling SMD simulation at T=310K of the DSB 4-bp.

ing across a crystalline plane in a solid (in which case, the sinusoidal peaks in the force-displacement plot correspond to the subsequent overcoming of identical Peierls barriers in the crystal lattice). To provide a visual representation of some key events during this complex sliding-bond-switching process, Figure 64 displays successive snapshots of a DSB 4-bp MD simulation, centred about the DSB region. The sugar-phosphate backbone is represented by continuous ribbons, clearly showing the sites of the two opposite strand breaks.

- In the upper panel (a), at t=0.6 ns, the H...O bond between G_{18}-C_{14} is shown to be cleaved (purple-dashed line, elongated beyond 4 Å-length); at this stage, all the other H-bonds between the base-pairs comprised in the 4-bp fragment (T_{19}-A_{13}, G_{18}-C_{14}, T_{17}-A_{15}, A_{16}-T_{16}) are still active, although stretched and bent.

- In (b), taken at t=5 ns, two H-bonds survive between G_{18}-C_{14}, however a non-W-C bond has formed between T_{19}-C_{14}, while T_{19} retains one regular H-bond with A_{13}. This is the signature of the beginning of the progressive "sliding" movement of the two broken strands, which will last until about t=34 ns.

- In (c), at t=31.8 ns, the two strands have largely shifted, and are kept together by the non-W-C bonds between T_{19}-C_{14} and between G_{18}-A_{15}, plus the π-stacking interactions between these four bases; the other bases in the DSB 4-bp fragment, namely

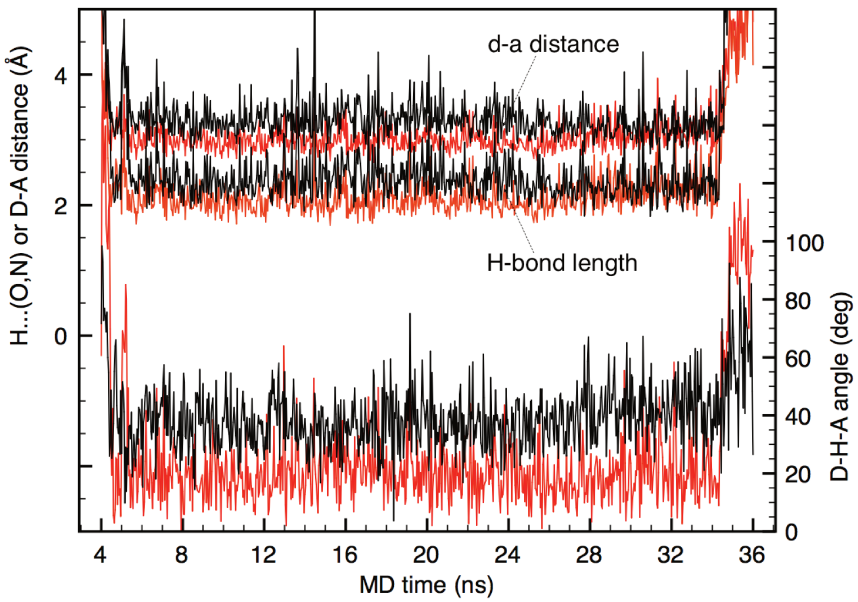


Figure 63: Time evolution of the donor-acceptor distance, H-bond length, and H-d-a angle, formed by $T_{19}H1$ with $A_{15}N3$ (black lines), and by $T_{19}H21$ with $A_{15}O4'$ (red lines), during a force-pulling SMD simulation at T=310K of the DSB 4-bp. D-A distances fluctuate about the average 3 Å, H-bond lengths between 1.8 and 2.2 Å. The H-d-a angle for the H1...N3 (black) fluctuates at the limit of our cut-off value of 30 deg, indicating a weaker bond.

A_{13} and T_{16} on one strand, T_{17} and A_{16} on the other strand, fluctuate sideways from the sugar-phosphate backbone. Also, note that under the twisting force, the whole 4-bp fragment has now turned by about 180 deg about the z-axis, as shown in Figure 65. These H-bonds are progressively deformed and brought to tension, up to the third peak at t=30-34 ns (about 32-36% elongation with a ~80 pN peak force in Fig. 60), after which they are cleaved, and the stress is released.

- Finally in (d), at t=35.7, we see the two G_{18}'s H atoms making up the two quite short, unusual H-bonds with two T_{16}'s O atoms from the backbone. Such a final sliding step corresponds to the last bonds (orange-dashed in Fig.61) setting in; these short-lived bonds have a length fluctuating around 1.8 Å, and are the only bonds (plus the G_{18}-T_{16} π-stacking interaction) to held the two strands together in the final part of the stretching, before arriving at the break-up of the two DNA fragments, at t ~37 ns. Here the tensile force in Fig. 60 drops to zero, with the fourth broad peak of ~30 pN height, center at ~40% elongation.

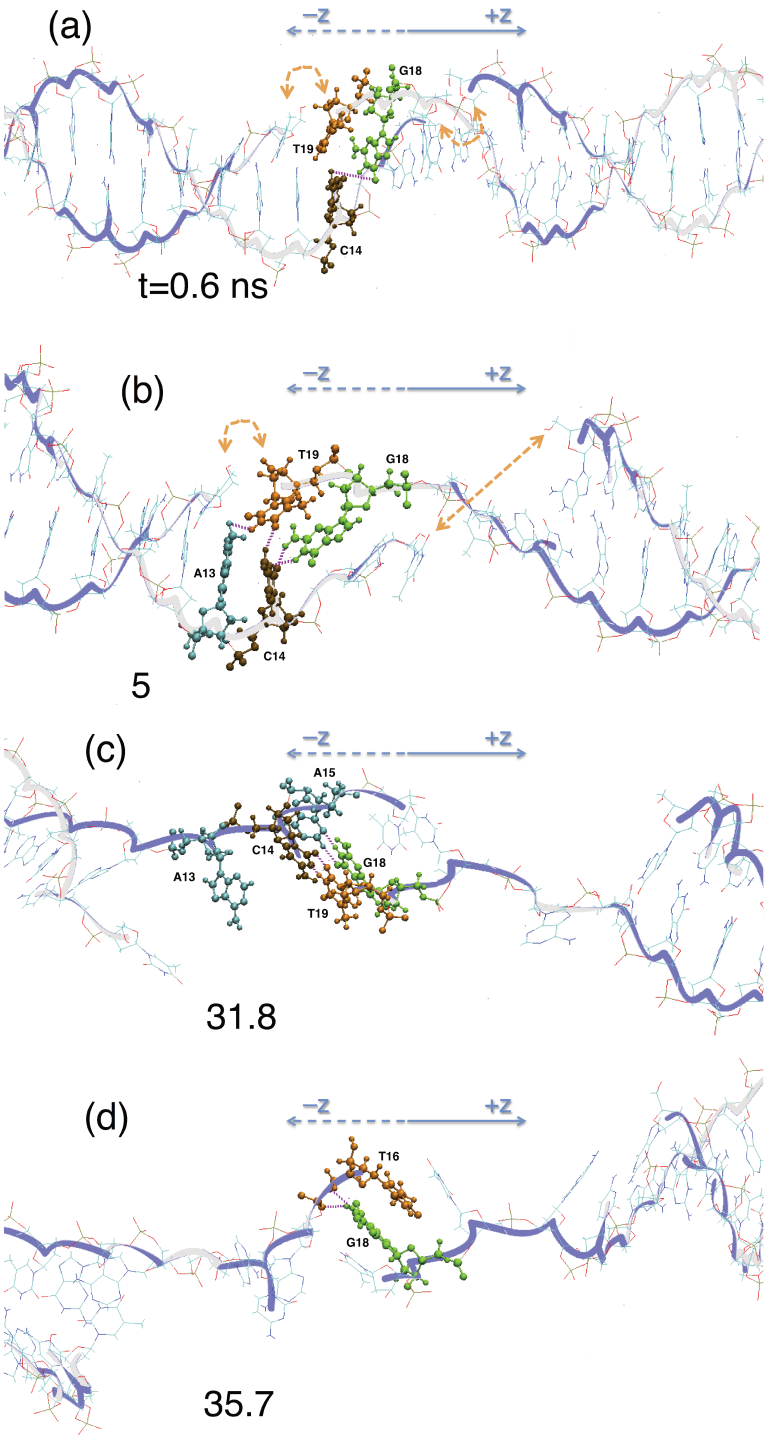


Figure 64: Snapshots from a MD simulation at T=310 K, for the DSB 4-bp; pulling direction along the main z-axis; fictitious spring constant 1,660 pN/nm, pulling velocity 4 cm/s; frame-time units 0.02 ns. Base color codes: A=cyan, C=burgundy, T=orange, G=green. Purple-dashed lines indicate H-bonds. The curved or straight orange-dashed orange arrows indicate the position of the two opposite strand breaks.

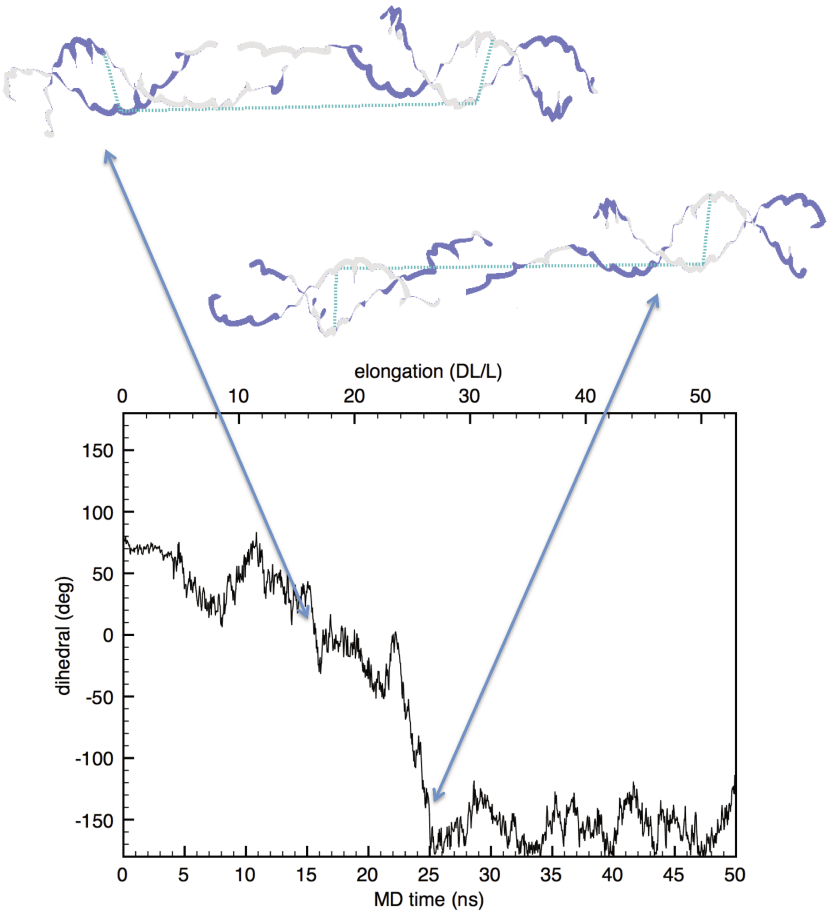


Figure 65: Time evolution of the fictitious dihedral angle formed by four C4′ carbon atoms from two bases on the left DNA fragment and two bases from the right DNA fragment (see the two DNA-ribbon scheme above the plot), during a force-pulling SMD simulation at T=310 K of the DSB 4-bp. The dihedral is reversed by about 180 deg between t=15 and 25 ns.

Coming back to the force-displacement curve, the first peak between about 16% and 25% elongation in Fig.60 is not directly linked to bond breaking, but rather to a global rise and release of stress, notably the twist stress of one half of the undamaged DNA with respect to the other half, with the DSB fragment acting as joint. This ample twisting movement is demonstrated in Figure 65. There we represent the fictitious dihedral angle formed by four C4′ atoms belonging to the ribose of four distant bases. In the interval t=15-25 ns, this dihedral switches from $\langle\Psi\rangle$=+40 to -150 deg.

The second peak in the force-displacement plot (centred at about 27% elongation in Fig. 60) is associated with the opening of the left-side strand-break, which occurs right after the above structure twisting is completed: at this stage, the maximum of the pulling force gets localized around this break region. The kinetics of this second collective process is shown in Figure 67, with a sequence of snapshots covering the interval t=14-27.5 ns (about 15 to 30% relative elongation). Moreover, at the site of the strand break a new H-bond forms

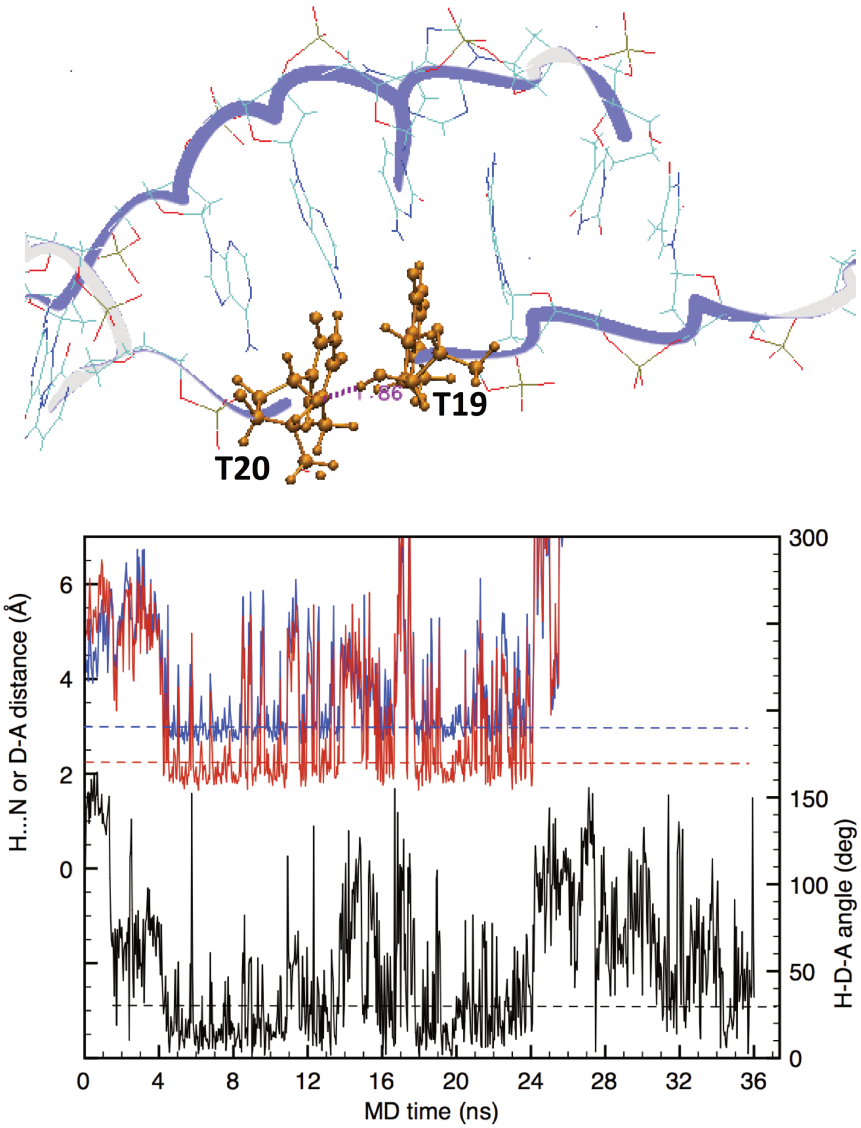


Figure 66: Time evolution of the T_{19}–OH...O4′T_{20} "unusual" H-bond. This forms at the strand-break site between times t $\simeq$4 ns and 24 ns (see ball-and-stick model above, with the H-bond indicated by the purple-dashed line), during a force-pulling SMD simulation at T=310 K of the DSB 4-bp. The blue plot is the donor-acceptor distance, with the blue-dashed line indicating d_{DA}= 3 Å. The red plot is the H-bond length, with the red-dashed line indicating the upper value 2.2 Å(both read on the left ordinates). The black plot is the H-d-a angle, the black-dashed line indicating the upper cut off of 30 deg (right ordinates).

around 4 ns and survives up to the strand opening at 25 ns: as shown in Figure 66, this new bond occurs between the terminal H atom of the 3′-phosphate termination of T_{19} and the O4′ of the T_{20} ribose. Such an unusual bonding configuration could be due to our particular choice of the 5-OH/3-PHO termination of the strand-break, and could happen differently for a different chemical termination; however, the reactivity of the ribose O4′ has been pointed out in several

studies [47], making the present observation one realistic outcome among other possibilities.

6.5 BREAKING BY THERMAL EXCITATION

The breaking of DSBs under an external force gave a wealth information about the detailed dynamics of the process. However, at this stage it is not possible to say nothing about the propensity for such breaking event to occur spontaneously under zero force, which could e.g. allow to formulate a ranking of the defects according to their lifetime. Of course, it could be easily deduced that the wider it is the DSB cut spacing, the more energy is needed to be overcome before arriving at the full break up of the molecule. However, such a qualitative deduction should be supported by a more quantitative assessment. Therefore, we tried to attribute a lifetime to the different DSBs by performing thermal annealing runs over a range of temperatures, aiming at extracting the prefactor from an Arrhenius-like plot. This endeavour is complicated by the fact that increasing the temperature above the physiological values, DNA should start denaturating; moreover, even if DNA should last enough time in the compact double-helix, there is no guarantee that the force fields fitted on low temperature properties could be reliable at too high temperatures. For this reason, we tried to run MD simulations within a relatively strict temperature range, and over run times for which the DNA does not show any signs of mechanical instability.

The thermal stability of the three DSB configurations was studied by constant-{NVT} MD runs of up to 200 ns (and 500 ns in just one case), preceded by a {NPT} thermalization run of about 2 ns, and carried out in a range of temperatures between 300 and 400 K (exceptionally up to 500 K), with repeated thermal annealing cycles between the lower and the target temperatures.

The force-pulling simulations provide a very complex picture of the evolution leading from a DSB to a fully-cleaved DNA. One key observation is that the integral under the curve is, as could be expected, roughly proportional to the number of H-bonds comprised within the DSB fragment. This integral is about 12 kcal/mol for the 1-bp, ~30 kcal/mol for the 2-bp, and ~100 kcal/mol for the 4-bp DSB configurations (depending on our particular choice of the DNA sequence in the DSB). Not only this is quite a large amount of energy: moreover, as shown by the force-displacement plots, the energy surface is characterized by high barriers, which require to apply a large steady force for a substantial amount of time, in order to break the damaged DNA. The big question posed by such simulation results with respect to *in vivo*, radiation-damaged DNA, is therefore: whence such forces and energy could arise? Possible origins can be found, alternatively or concurrently, in:

(i) internal tension from the chromatin structure;

(ii) external action from specialized proteins;

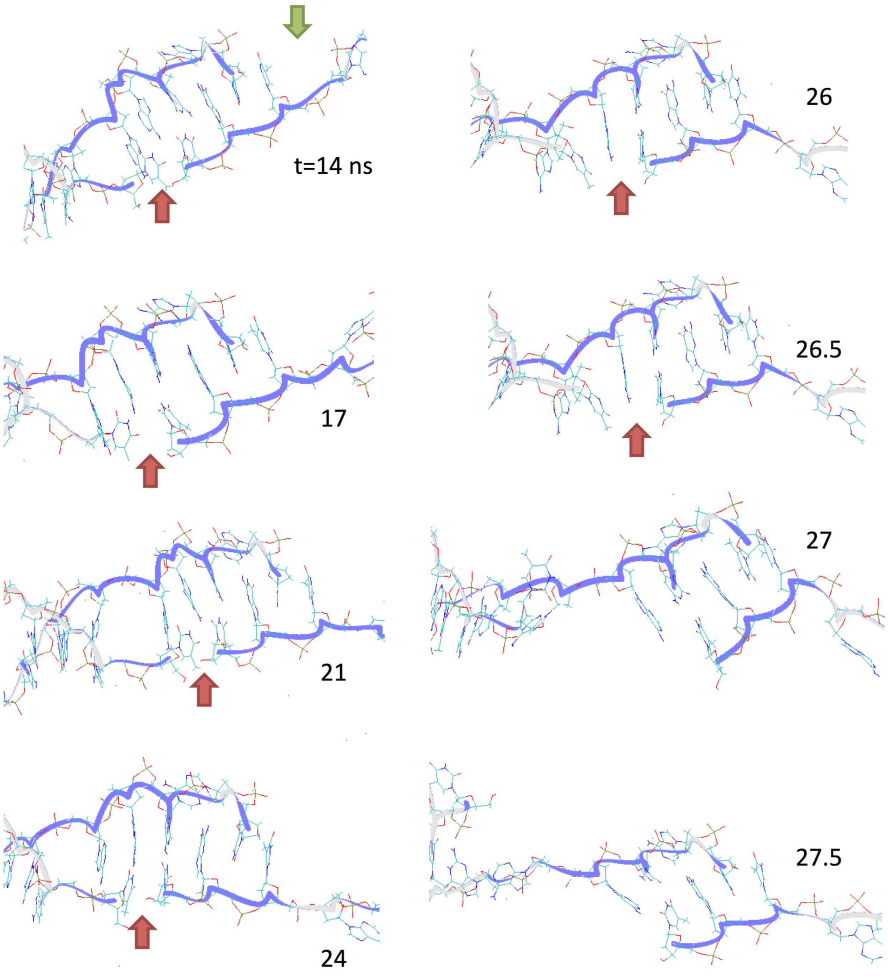


Figure 67: Time sequence of eight snapshots of the DSB 4-bp region from a force-pulling SMD simulation at T=310K. The right-side strand break (green arrow) is already open before the first time-frame shown here, at t=14 ns. The left-side strand break (red arrow) is still closed, until the final opening occurring around t 26 ns. After this time, the three bases of the lower strand continue to slide to the right with respect to the four bases of the upper strand, which slide to the left of the figure. The fourth base of the lower strand can be seen at the lower right corner, swinging free to the back of the sugar-phosphate backbone, already from the early stages of the pulling simulation.

(iii) thermal fluctuations.

Concerning (i) internal chromatin forces, the available data from single-molecule pulling experiments on *in vitro* reconstituted chromatin fibers or nucleosome arrays found two pulling regimes [36, 121]: a low-force regime, exhibiting a force plateau at ~5 pN, and a higher-force regime exhibiting saw-tooth patterns. These experiments indicated that, for forces between 6 and 20 pN, chromatin underwent stretching without any structural transition. On the other hand, stretching of the fibre around and beyond 25 pN resulted in the release of the histone octamers [21] . The external forces imposed onto chromatin by DNA-based molecular motors, such as RNA and DNA polymerases [40, 52], and ATP-dependent chromatin remodelers [24, 95], typically fall in the range of a few tens of pN; RNA polymerases have been shown to be capable of exerting peak forces as large as 40 pN. In all cases, such transient forces seem to be below the values of ~80-100 pN we observed in our simulations. While some uncertainty can be attached to the force constants of the chosen molecular force field (CHARMM-27), such a large difference by a factor of 3 to 4 seems to make damaged-DNA breaking by purely internal chromatin forces quite unlikely.

The (ii), external action from specialized proteins, may be actually operating in some cases, such as enzymatic cleavage [75], or after arrest of the replication fork [70]. Also, it is established that in the middle stage of DSB repair, specialized proteins, such as Artemis, intervene to clean up the hanging ends of the strand cuts [169]. However, all such cutting enzymes are unlikely to be effective in the early stages of DSB evolution, since at this time damage-signaling proteins are rather the ones active around the damage site, such as PARP in SSB [133], and the MRN complex in DSB [87].

Thermal fluctuations (iii) have been invoked as a possible cause of the eventual break up of the DNA containing a distribution of SSB and DSB [162]. The energy barriers obtained in the force-pulling simulations appear to make thermal disruption unlikely, unless the DSB is very short-cut. However, it must be noted that the SMD algorithm puts a strong bias on the reaction coordinate, and it cannot be excluded that a more thorough exploration of the phase space could lead to more favourable configurations for breaking. Therefore, we performed a series of thermal stability simulations for the three DSB 1-bp, 2-bp and 4-bp structures, at temperatures ranging from 300 to 400 K (or 500 K for the DSB 1-bp, going to even higher temperatures would not make sense, since the fitting of the force field becomes totally unreliable). The simulations were initiated by a ~10 ns run at low temperature and constant-{NPT}, followed by long runs at constant-{NVT} conditions. Rigid TIP3P water molecules filled a parallelepiped box of 5x5, or 6x6 nm^2 cross section, and length ~15 nm, leaving enough room for the damaged DNA to experience ample fluctuations also in the directions non-parallel to its main axis. For the higher temperatures, the volume in the constant-{NVT} was fixed for

values of pressure between 50 and 300 MPa, corresponding to water always in the liquid state.

All such runs were performed, as above, with positional restraints applied at the two ends, to simulate the chromatin background of the linker-DNA. The DNA ends were constrained by two soft springs, allowing a fluctuation in x, y and z of about ± 1 Å RMS. However, because of temperature excitation it also was necessary to apply additional soft harmonic constraints between the $5'/3'$ ends, to avoid opening of the helix after very long simulation times. These latter constraints are quite far away from the DSB central region, and did not influence its local dynamics.

The DSB 1-bp configuration is clearly the weakest, since it is held together only by the two H-bonds between the A_{16}-T_{16} bases. Therefore, its rupture is directly observable even in relatively short MD simulations, with a rapid acceleration upon increasing the temperature above ambient (this makes it possible to raise the temperature up to 500 K, since the rupture is fast enough to avoid thermal denaturation). However, contrary to a simplistic expectation, the rupture kinetics seems dominated by the π-stacking interactions, and not by the cleavage of the H-bonds. This effect can be appreciated in Figure 68, a time sequence of six snapshots of the DSB at T=350K, and in Figure 69(a), which shows the evolution of the vertical distance between the aromatic cycles for the A_{16}-T_{16}, A_{16}-A_{15}, and A_{16}-T_{17} bases for the same simulation.

The two A_{16}-T_{16} H-bonds are broken within the first ns of the simulation; at the same time, a temporary H-bond is formed between the A_{16}OH and the O4′ of the T_{17}, similar to what observed for the force pulling of the 4-bp configuration. After this switching of H-bonds, the two bases shift to a stacked configuration (see also Fig. 7 Suppl. Data at t=10 ns), in which the π-orbitals can provide sufficient binding to maintain the DSB still closed. The black trace in Fig. 69 represents the vertical distance between the centers of A_{16} and T_{16}, showing that this configuration persists up to t $\simeq$15 ns, and reappears at t $\simeq$35 until the final break-up at t=55.5 ns. In the time span between 15 and 35 ns, the T_{16} temporarily rotates by about 180 deg about the sugar-phosphate backbone, thus disrupting its π-stacking interaction, which is superseded by the A_{16}-A_{15} (red trace) and the A_{15}-T_{17} (blue trace) stacking interactions.

In this DSB 1-bp case, the bonding lifetime at the various temperatures can still be fitted by a Bell-like equation [17], representing the dissociation of a single, "lumped" system comprising both the hydrogen bonds and stacking interactions:

$$\tau = \tau_0 \exp(E_b / k_B T) \tag{6.1}$$

The results are reported in Figure 69(b), from which the time constant τ_0=2.5 ps (representing the time to rupture at very high temperatures), and the binding energy E_b=7 kcal/mol, can be deduced. (Note that the first constant is affected by a rather large error, since it comes from the extrapolation over several decades in the semi-log plot.) The value of E_b is about 40% smaller than the value of 12

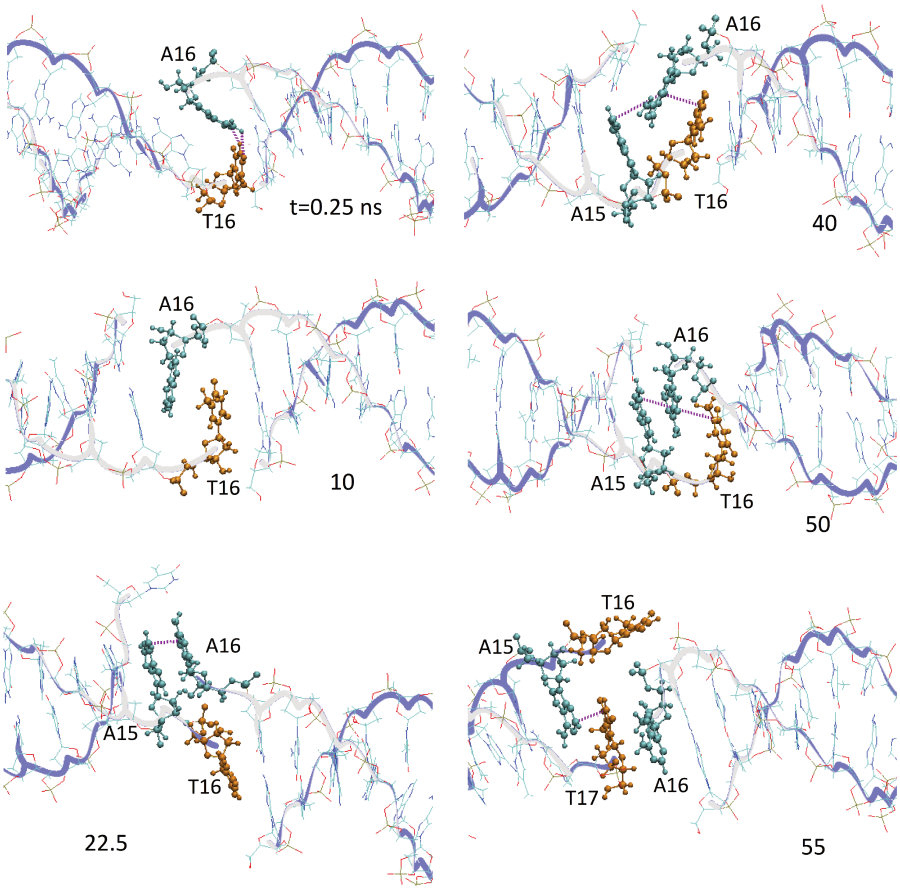


Figure 68: Time sequence of six snapshots of the DSB 1-bp region, from a unconstrained MD simulation at T=350K. The H-bonds between A_{16} and T_{16} are broken within 1 ns from the beginning. At t=10, the two bases are shown in the new stacked configuration. At t=22.5, the T_{16} has rotated by 180 deg about the backbone, and the new stacking interaction takes place between A_{16}-A_{15}. At t=40, the T_{16} has rotated back into place, but the DSB configuration is much distorted. At t=50 ns the DSB appears isolated from the rest of the DNA. At t=55 ns, immediately before the final break-up, the T_{16} and A_{16} are disconnected, and the residual A_{16}-T_{17} stacking interaction is the last one to be split apart.

kcal/mol deduced from the integral under the curve of the force-displacement plot (Fig. 60, top curve). This difference is likely due to the different bonding structures of the DSB in the two simulations. In fact, the force-pulling SMD simulation is highly constrained along a unidimensional reaction coordinate, moreover at a very high pulling speed compared to real experiments, whereas the thermal rupture occurs spontaneously with the DNA freely exploring a much larger configurational space. In particular, the complex sequence of stacking interactions of Fig. 69 and Fig.68, is not observed in the SMD simulations. Therefore, the 7 kcal/mol can be mostly attributed to the energy barrier of the π-stacking interactions, while in the SMD the value of 12 kcal/mol rather represents the barrier from the H-bonding interactions.

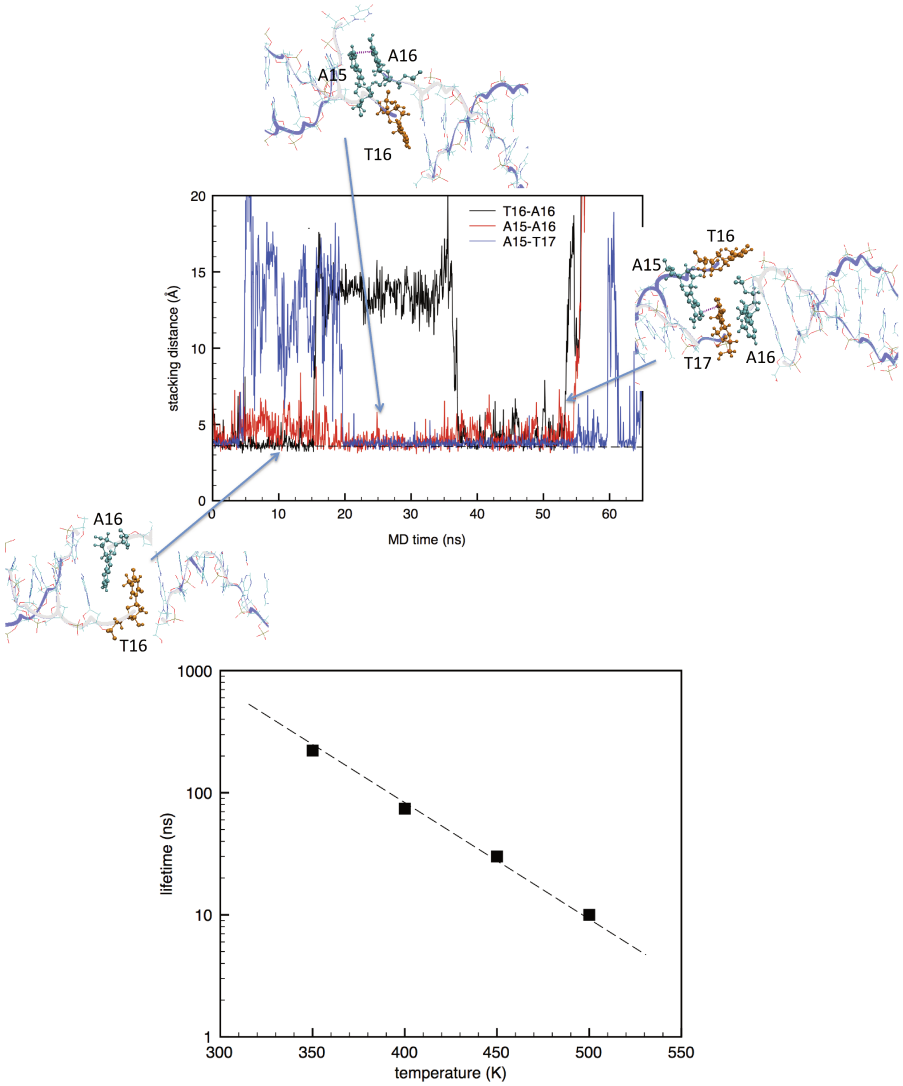


Figure 69: **(a)** Time evolution of the vertical stacking distance between the base-pairs indicated in the legend (black=A_{16}-T_{16}, red=A_{16}-A_{15}, blue=A_{16}-T_{17}). The dashed line indicates the average stacking distance of 3.4 Å. The final rupture event occurs at t $\simeq$55.5 ns, indicated by the sudden rise of the black/red distances. Snapshots from the constant-{NPT} MD simulation at T=350 K indicate the typical stacking configurations at t = 10 ns (lower left), 22 ns (above), and 50ns (right), just before the final rupture. **(b)** Fit of the observed bonding lifetime for the DSB 1-bp as a function of the simulation temperature. Full squares represent MD data.

Directly observing DSB breaking at larger strand cut distances was very difficult. Already for the DSB 2-bp the time to rupture grows rapidly. In practice, we could observe a breaking within the first few tens of ns only when the initial conditions were (on purpose) extremely strained, and notably for unconstrained DNA. Under "linker" initial conditions, namely a nearly straight DSB 2-bp, with positional restraints at the two ends, pre-relaxed at low temperature for more than 10 ns in order to allow a proper distribution of water and counterions, breaking was not observable up to 200 ns for T $\leqslant$400 K (at higher temperatures, our DNA model denaturates after just a few ns). More-

over, not only is damaged-DNA resistant to thermal disruption, it is also *resilient*. In one simulation at T=400 K and P=270 MPa with unconstrained ends, between t=40 and 45 ns the DSB was observed to open up to a quite extreme configuration, being held in contact only by 3 residual stacking interactions and a few, loose H-bonds; however, in the span of about 5 ns it closed back to the initial configuration, with all the H-bonds and base stacking properly rearranged, before finally breaking at ~58 ns, when the periodic boundary conditions forced the unconstrained DNA fragment to start interacting with its periodic images and the simulation becomes meaningless. We also tried cycling up and down between 350 and 400 K the same configuration without observing destabilization of the DSB. Our longest simulation for the DSB 2-bp was of 300 ns; while such a simulation time is still far below the longest DNA simulations published to date [55], it is worth noting that our DNA duplex are quite longer.

The above findings are even more reinforced for the DSB 4-bp, for which we could never observe thermal disruption at $t \leqslant 200$ ns, under none of the tested conditions at $T \leqslant 400$ K. Clearly, due to exponential nature of the bond lifetime, the actual rupture falls beyond the possibility of MD simulations, and even increasing by a factor of 10 the simulation time, we could hardly hope to directly observe the rupture event.

If we apply the Bell-like equation to the situation of m bonds in parallel, a gross estimate of the bonding lifetime can be obtained, by simply multiplying by m both the τ_0 and E_b parameters deduced from the 1-bp study. In this approximation, the extrapolated values of lifetime for the DSB 2-bp at 310 K would be of the order of 50-100 milliseconds, and for a hypothetic DSB 3-bp this value jumps to $\tau \simeq$ 1-2 hours. Even conceding a large error bar to such estimates, it can be seen that DSBs should be highly stable in temperature is substantially correct. We can suggest that only very close-cut DSBs could actually experience spontaneously the complete DNA break-up, whereas if the two strand breaks lie at a distance of just 3 base pairs or more, the DSB is practically immune to thermal disruption. If confirmed, such findings could bear a huge impact also on the dynamics of the repair proteins.

6.6 FINAL SUMMARY

In this Chapter we studied the evolution under external force and temperature of radiation-induced strand breaks in a linker-DNA fragment, with end-constraints to represent the embedding in the chromatin structure. We investigated the possibility of identifying by MD simulations, mechanical signatures or "fingerprints" that could help the experiments discriminating between different microscopic configurations of strand breaks. While vibrational spectra analyses do not seem to contain the right (or, the right amount of) information, force spectroscopy appears, instead, to reveal a rich dynamics, allowing not only to distinguish between different types of DNA damage, but also

to learn a lot of important microscopic details. From our simulations, the absolute values of force necessary to break up a DSB-damaged DNA are very large, of the order of 100 pN, at elongations above ~20% (notwithstanding the approximations implicit in the use of empirical force fields). Such values of longitudinal stress and strain are unlikely to be observed in the normal dynamics of chromatin, or during chromosome mitosis. By comparison, thermal fluctuations seem unable to provide the energy necessary to overcome the barrier to rupture, unless the DSB is a very close-cut one (i.e., the two breaks on opposite strands are separated by up to 2-3 base pairs at maximum).

A detailed knowledge of the structural and mechanical response of DNA after radiation-induced damage is very relevant, since the repair machinery has a very high sensitivity to the strand-break position and conformation, besides dependence on the cell cycle phase (as pointed out in Section 1.3). A key question is therefore: "what" the scouting proteins actually recognise at the damage site? For example, it has been postulated that, in the early stages of NHEJ, the Ku70/Ku80 heterodimer firstly binds to the open ends of cleaved DNA, on the basis of x-ray structures in which DNA fragments are co-crystallised with the monomers [127, 144]. However, this should be true only if we admit the DNA is firstly completely cleaved into physically separate fragments. On the other hand, we showed that even after the DNA backbone has been cut on both sides, by either a direct or indirect action, a considerable binding energy from the non-covalent interactions still remains, to keep the fragments together (hydrogen bonds between the nucleotides, stacking interactions among the vertically-piled aromatic cycles, electrostatic screening by ions). If one instead postulates, on the basis of the results of our MD simulations above, that DNA may not be fully broken, even after substantial radiation damage, could it be possible that such proteins have an even stronger affinity for some of the intermediate states, as shown e.g. in Figure 1.3? Could such proteins be able to identify *severely damaged*, rather than fully cleaved DNA, and what the implications could then be, for the subsequent steps of the repair chain? OR maybe proteins like Ku70/Ku80 identify *only* the broken DSBs, and completely ignore the others?

An important implication of our findings is that DSBs actually undergoing spontaneous breaking by thermal fluctuations are only those with a short strand-break separation. Those DSBs in which the strand breaks are 3-4, or more base-pairs apart can deform, indeed, and give rise to transient extremely distorted configurations; however, the DNA strands seems to retain mechanical connection and resistance. Our simulations demonstrate that the π-stacking interaction can be strong enough to take the place of broken hydrogen bonds, in holding together the severely damaged DNA strands. Under external tension and torsion forces, instead, shorter DSBs split quite efficiently, after the few hydrogen bonds holding the bases together are cleaved. On the other hand, the breaking of larger DSB proceeds by complex "stick/slip" sliding mechanisms, yet requiring substantially large forces and deformations. For the sake of completeness, it

is worth noting that we did not consider at this stage of our study the possible clustering of DSBs, or their association with other types of defects, e.g. abasic sites, into more complex lesions.

On the basis of such observations, it is tempting to formulate a final hypothesis for this part of the study. In the whole spectrum of strand-breaks created by a given dose of ionising radiations, there could be different populations, corresponding to different separations between the opposite strand-breaks: among these, only the closest DSBs (distance $\leqslant$ 2-3 bp) may lead to complete fracture ("active" DSBs). If unrepaired, these are the defects actually leading to cell arrest and chromosome aberration. Moreover, if our hypothesis is true, the number of DSBs detected in an experiment must depend on the technique used: if proteins sensitive to the open strands are tracked by fluorescence, a value close to the total number of DSB would be counted; if, on the other hand, a technique tracking only broken DNA fragments is used, such as the comet assay, only the "active" DSBs should be observable. The possible impact of such profound differences also on, e.g., calibration of dose-response curves and tumour-control probability, to establish the efficiency of radiotherapy protocols, are certainly open to further investigation.

7

NUCLEOSOMAL DNA SIMULATIONS

The region of freely exposed (linker) DNA represents only a minority fraction of the total length of the DNA in the chromatin. As sketched in the Introduction, higher levels of spatial organization are adopted to compact and protect the genomic information, from the 10-nm to 30-nm fiber to chromosomes. Anyway the "active region" of the chromatin in the nucleus presents usually lower levels of organization, in order to maintain an easy accessibility to the DNA sequence for the *transcription* process. As a consequence the nucleosome, the building block of chromatin, and linker DNA are the basics units needed to understand the DNA behaviour in active regions.

In this perspective, the presence of a strand break in the"active area" of DNA is more likely to cause irreversible damage to the cell's life. Using the knowledge accumulated in the study of the linker DNA, it was decided to use MD simulations also to observe the mechanical evolution of the most lethal defect, a short-cut DSB, within the DNA portion coiled around a histone core. As already said, DSBs in the DNA backbone are the most lethal type of defect induced in the cell nucleus by chemical and radiation treatments of cancer, and are the ones that are likely to induce the most significant deviation of mechanical properties compared to the undamaged sequence. Little is known to date about specific outcomes of damage in nucleosomal DNA, and on its effects on the damage repair cycle. Taking the configuration of the DSB with one base-pair width from the previous study, we introduced this type of defect at different positions in the nucleosomal DNA, and analysed the behavior of this defect.

7.1 MOLECULAR STRUCTURES OF THE DAMAGED NUCLEOSOME

We obtained the nucleosome molecular configuration from the RSCB Protein Database, PDB-entry 1kx5 [41]. This is an x-ray structure of the entire histone octamer with 147 DNA bp resolved at an average RMS of 1.94 Å, reconstituted from human nuclear extract expressed in *E. coli*; only 6 histone residues were unidentified in this experimental structure, with respect to the known histone sequences, therefore the model can be considered nearly complete. The 147 bp DNA is a palindromic sequence, chosen to maximize the degree of ordering and increase the x-ray spatial resolution. To obtain a model structure useful for our computer simulations, we removed all the crystallization water molecules and ions from the published structure, and added two DNA extensions of length 20 bp at each end of the nucleosomal DNA, with repeated sequence d(AGTC) [137].

DNA bases are numbered from 1 to 187 in each chain, one running clockwise and the other counter-clockwise, the dyad being located at basis 94 of each chain. This pristine nucleosome model without

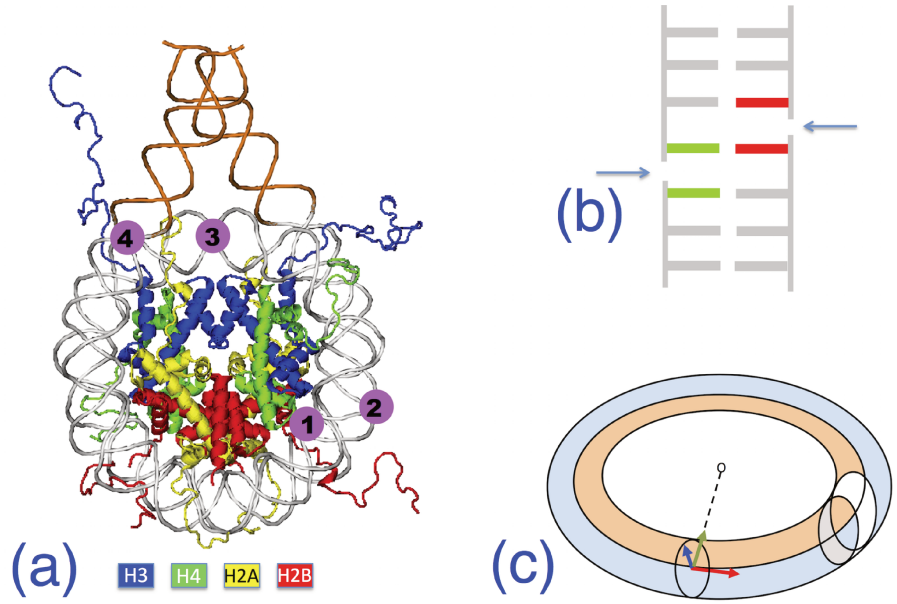


Figure 70: **Nucleosome and DNA defect geometry. (a)** Schematic of the experimental nucleosome structure 1kx5, with DNA in grey; in orange, the 20 bp added to the experimental structure on each end. Histone pairs H3, H4, H2A, H2B color-coded as shown below. Positions 1 to 4 indicate insertion sites of double-strand breaks (DSB). **(b)** Schematic structure of a DSB; arrows indicate the cut on each backbone; the green-red central base pair is an A···T for all models. **(c)** The circular bent-tube DNA geometry in the nucleosome. At any point along the neutral axis (centerline) a local reference frame is defined by three unit vectors: normal **n** (blue), tangent $\boldsymbol{\tau}$ (red), and the binormal **b** (green) directed to the center of curvature. The white slice represents a tube averaging zone for stress calculations.

strand breaks is shown in Figure 70(a), and will be labelled O in the foregoing.

DNA is wrapped left-handed about the histone core, making two nearly complete turns that join at the dyad symmetry point; the two DNA turns define two circles lying in two ideally parallel planes, with a superhelical symmetry axis perpendicular to the center of the circles (for a thorough discussion of nucleosome geometry and structure, see e.g. Ref. [107]). The relaxed DNA double helix makes a complete twist around its double-helical axis, about every 10.4 bp, defining a major and a minor groove; therefore, when turning around the histone core, the wrapped DNA makes 14 nearly full twists. Correspondingly, 14 contact points between DNA and proteins can be identified within the nucleosome structure, loosely situated at the minor groove locations facing inwards.

Based on these geometrical features, we defined 4 potentially interesting sites along this wrapped structure, where to place a DSB in a "mechanically significant" position labelled 1 to 4 in Fig 70(a). Correspondingly, we introduced a DSB model for each position[1]:

1 The – symbol indicates the break site along each backbone, the ··· indicate the central interacting base pair.

- *model M1*: an inner contact site, between bases C69-T68$\cdots$A120-G121 ;
- *model M2*: an outer non-contact site bases C73-A74$\cdots$T114-T113;
- *model M3*: at the dyad, between bases A94-T95$\cdots$A94-T95 of both chains;
- *model M4*: at the entry point of the nucleosome, between bases T22-A21$\cdots$T167-G168.

The DSB is described, as before, by introducing the 5′-OH and 3′-phosphate terminations at each end of the cleaved strands. In this way, the two backbone cuts of each DSB are spaced by 1 bp always comprising an A$\cdots$T pair (Fig 70(b)). The DSB is initially bonded by only its two hydrogen bonds, plus the stacking interactions on each intact side of the chain, while the other half of stacking starts to get readily reduced, as soon as the MD relaxation starts.

7.1.1 *DSB dynamics at different nucleosome positions*

For nucleosome MD simulations we used only the GROMACS 5.1 computer code [18, 92], because we observed a considerably better efficiency compared to NAMD, for the system sizes necessary in this case. Nucleosome models O and M1-M4 were solvated in water box of size 14.5 or $18\times19\times10$ nm^3 with periodic boundary conditions in the three directions, containing about 82,600 or 110,500 TIP3P water molecules, plus 480 Na^+ and 250 Cl^- ions to ensure neutralization of the phosphate backbone charge, and a physiological salt concentration around 0.15 M.

Following our established protocol, long preparatory annealing cycles of the water and ion background, while keeping the nucleosome still, to obtain the right water density and allow a realistic arrangement of the counter-ions around the phosphate backbone, prior to starting the microsecond production runs.

All the MD simulations were carried out at the temperature of 310 K and pressure of 1 atm, or 350 K and 50 atm for the thermal stability study (at constant-{NVT}, hence the small overpressure within the typical numerical fluctuation for a system of this size, corresponding to the experimental pressure of water at 350 K).

The DNA terminal ends, which represent a portion of the DNA *linker* in real cells, were restrained by soft harmonic constraints, allowing a fluctuation of ±5 Å, to represent embedding in the chromatin structure. We used rigid bonds for the water molecules, which allowed to push the time step to 2 fs for the thermal equilibration runs, and to 1 fs for the force-pulling simulations. Typical preparatory constant-{NPT} MD runs lasted between 10 and 20 ns; force-pulling simulations were carried out for 10 ns, and the subsequent force-free relaxation lasted up to 400 ns; thermal stability simulations at constant-{NVT} extended to ~1,000 ns for O and M2-M4, and up to 1,800 ns for the M1 model. Overall, this part of the study used about

Table 7: List of the main MD simulation trajectories, deposited in condensed format in the public repository Figshare, doi: 10.6084/m9.figshare.5840706. All trajectories in this list are simulated with constant-NVT (the initial equilibration at constant-NPT is omitted). The numerous force-pulling trajectories are not listed.

Label	DSB configuration	Temperature (K)	Duratio (ns)
M1-HT	M1	350	1800
M1-LT	M1	310	500
M1-force-rel	M1	310	100
M2-HT	M2	350	1020
M2-LT	M2	310	300
M3-HT	M3	350	980
M3-LT	M3	310	300
M4-HT	M4	350	980
M4-LT	M4	310	300
O-HT	O	350	500
O-LT	O	310	500

4,2 million hours of CPU time on 2048 IBM BlueGeneQ processors (IDRIS supercomputing center in Orsay), and about 800,000 hours on 896/1064 Broadwell Intel E5-2690 multi-core processors (CINES supercomputing center in Montpellier), with typical running times of 1.3 and 7 ns/hour on the IBM and Intel machine, respectively. About 1.5 Terabytes of raw data were accumulated over a period of 8 months, from March to October 2017, for subsequent post-processing (see Table 7).

Because of the requirements of the forthcoming stress calculations (see below), we could not use standard Ewald-sum electrostatics but were forced to adopt plain cut-off Coulomb forces. This is known to be at the origin of possible artifacts, therefore we used for both electrostatics and long-range non-bonding forces an unusually large cut off radius of 1.6 nm. Therefore, we ran some segments of trajectory with PME, restarting from previous configurations, and looked at some quantities to see whether there could be substantial differences between the trajectories so generated; in particular, we look at the differences for DNA, given its large negative charge; eventual differences should be even minor for the histone protein moieties, which have very small overall charges. For the M1 trajectory of 1,800 ns, we restarted from the configuration at time 980 ns, and ran 20 ns of trajectory with PME; the two segments of trajectory of 20 ns, with cut-off and PME, were then compared, by superposing the structures frame by frame. In figure 71, the RMSD between the two trajectories for the DNA, averaged base-by-base (20 to 167 for each strand), are reported at a few representative times. It can be seen that the RMSD remains in general well below 3 A (dashed line), with minor exceptions which however very rarely surpass 5 A. While it cannot be excluded that over much longer simulation times the two trajectories could even-

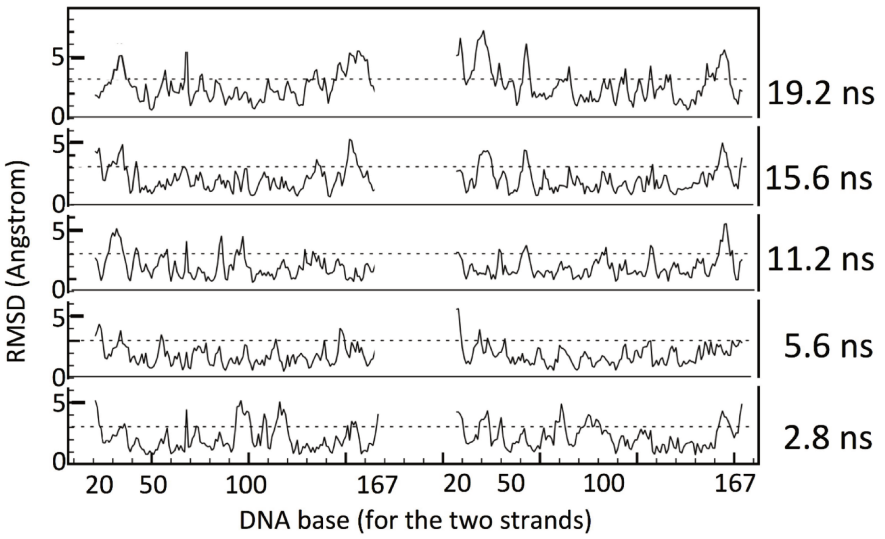


Figure 71: **Comparison of MD trajectory with PME vs. cut-off electrostatics.** For the M1 trajectory of 1,800 ns, we restarted from the configuration at time 980 ns, and ran 20 ns of trajectory with PME electrostatics; the two parallel segments of trajectory of 20 ns, with cut-off and PME, were then compared, by superposing the structures frame by frame. The plots show the RMSD between the two trajectories for the DNA wrapped in the nucleosome, averaged base-by-base (numbered 20 to 167 for each strand 50-30 and 30-50), at a few representative times. The dashed lines in each ploindicate the reference value of 3 Å.

tually differ, this test warrants a good similarity between trajectories obtained by the two different calculations of electrostatics.

We introduced in all models M1-M4 one single-bp DSB with a central A···T, which is the weakest bonded bp. Most MD simulations were run at the temperature of T=350 K, or about 77°C, in order to stimulate the thermal dynamics of the system, while remaining within a range of vibrational excitations that is still meaningful for the molecular force-field used. As seen in Tab. 7, MD trajectories extended to ~1 μs for the M2-M4 models, and up to 1.8 μs for the M1, since this latter displayed some potentially more interesting dynamical features. The reference model O with the intact nucleosome was simulated over a shorter trajectory of 500 ns. Shorter MD trajectories were also run at T=310 K for all models, for comparison.

We firstly present the results for the models M2-M4. For all three, we could not observe any substantial evolution of the DSB into a fully broken DNA, over the whole duration of the simulation, despite the relatively high temperature. While it cannot be excluded that such an event could be produced over longer times, this is an increase of more than a factor of 20 in lifetime compared to the linker DNA presented in the previous Chapter. Upon scaling by the same time and energy factors at 310 K, the corresponding dissociation time from the Bell-like model of Eq.(6.1) is estimated in the 100-μs time scale or longer, even for the most favorable (i.e., least bound) DSB configuration; this represents therefore a lower bound for the spontaneous dissociation

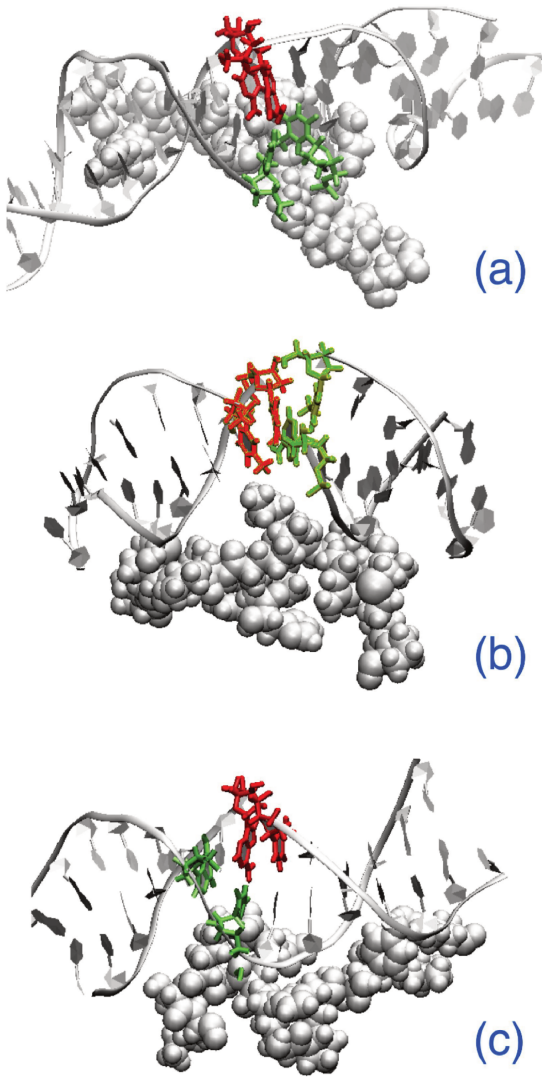


Figure 72: **Snapshots of MD simulations of the M2-M4 models of DSB**, after 1 μs of dynamics atT=350 K. **(a)** Model M2, with DSB at the outer non-contact site, showing the central A74···T114 bp still well bonded. Grey spheres represent a portion of the H3 histone flanking the defect. **(b)** Model M3, with DSB at the dyad. Grey spheres represent a portion of the H3 tail. **(c)** Model M4, with DSB at the entry point of nucleosomal DNA. Grey spheres represent a portion of the H3 tail close to the break, which has folded into a double α-helix.

time. Representative snapshots from the trajectories at the DSB sites are shown in Figure 72.

The bonding configuration of the central base-pair remains on average rather close to that of the pristine nucleosome, with the H-bonds providing a large fraction of the cohesive energy, and the mildly deformed stacking ensuring a substantial structure stability. An example can be observed in Fig. 73, in which the time evolution of the H-bond lengths for the central A···T bp of model M3 are shown. The three bonds formed by the N1 (adenine), O1 and O2 (thymine)

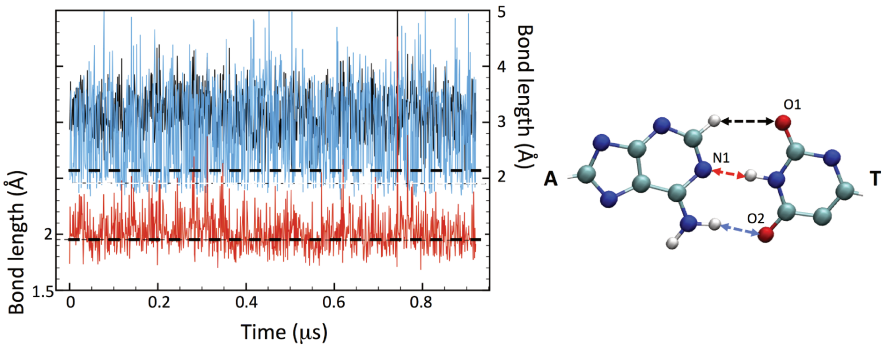


Figure 73: **Hydrogen bond length at a DSB base pair.** Plot of the H-bond length for the central A···T bp of the model M3, for a MD simulation of 950 ns at T=350 K. Red, blue and black traces relative to the respective arrows in the scheme above, with the A···T bp in ball-stick representation. The red arrow indicates the central H-bond with adenine N1, the blue arrow the side H-bond with thymine O1, and the black arrow the (very weak) H-bond with thymine O2. Horizontal dashed lines in the plot indicate the reference, T=350K H-bond distances of 1.9 (lower) and 2.1 Å (upper).

donors are indicated in red, blue and black, respectively. The relative strength of individual H-bonds in the A···T bp can be theoretically estimated [153] to be about 10:4:1 for the N1:O1:O2. The last one is not usually accounted as a true H-bond, since it is very weak and with a length fluctuating around 2.8 Å. Indeed, the central N1 bond remains always in the range [1.8-2.1] Å RMS (note that the simulation is at high temperature); the side O1 is more dynamic than the corresponding bond in normal DNA, with an average length of 2.3 Å (~2.05 in normal DNA), and quite large RMS fluctuations due to the larger rotational freedom of the DSB about the central axis; the O2 length remains well beyond the definition of H-bond, fluctuating about an average of 3.2 Å. Overall, these interactions provide enough bonding to keep the DSB in place, even in this M3-dyad position that is the farthest from the histone protein core, among all the DSB configurations studied.

7.2 ESSENTIAL DYNAMICS

To characterize the dynamic motion and capture anharmonic movements (bending, torsion, etc.) that dominate the global molecular dynamics of the DNA and of the closest protein residues around the DSB region, we performed a study of the essential dynamics (ED) for each model.

We firstly run the ED for the regions surrounding each location M1-M4 in the pristine nucleosome, model O. Typically, the analysis is restricted to a length of about 7 DNA bp on each side of the DSB, plus the 15-20 histone residues in the closest neighborhood of the DSB. MD trajectories are sampled at a rate of 40 ps^{-1}. Such analysis of the undamaged system provides a spectrum of eigenvalues, from which we extract the first few significant ones, and an average ref-

erence configuration for each M1-M4 site. Then, we repeat the same analysis on each of the independent trajectories including a DSB at the M1-M4 positions, by using as reference molecular structure the corresponding average from model O, so as to highlight deviations from the "normal" DNA dynamics.

A key quantity providing information about the large-scale movements of the fragments implicated in the DSB comes from the study of the first few eigenvectors, and of their root-mean-squared fluctuation (RMSF) on a atom-by-atom basis. Note that, like the RMSD, the RMSF is in principle measured in Å; however, being obtained from the eigenvector analysis, these *are not* actual atomic displacements, but components of a theoretical displacement projected out according to a particular deformation eigenvalue. Therefore we indicate the units as arbitrary, although they are numerically coincident with Å.

These new atomic variables capture the contribution of each group of atoms to the principal collective movements, as filtered out by the most important eigenvectors. For all the M1-M4 models, the first 4 eigenvectors are found to cover 65% of the weight, the 5-15 ones are responsible for another 20%, and all the remaining 3N-15 for the last ~15%. Such a distribution is less extreme for the O model, in which large-scale movements are quite more restricted, with the first 15 eigenvalues carrying about 55% of the total weight.

The physical meaning of such principal eigenvectors can be appreciated with the representative plots of Figures 74 and 75, in which the extreme configurations spanned by the large-scale motion of the first few eigenvectors are represented. For example, Fig. 74 displays the first eigenvector for the DNA fragments in models M1 and M3; all the frames, simultaneously represented, are colored from blue to red, the ordering showing how each atom's motion spans between the extreme values of the eigenvector. It can be seen that the principal eigenvector for M3 describes quite homogeneous, local fluctuations of all DNA bases, with just a more evident oscillation along the stacking direction concentrated about the DSB; on the contrary, for the M1 this principal eigenvector describes a dramatic large-scale displacement of the central atoms making up the DSB, which tend to span ample areas across orthogonal planes, by turning about the backbone. This largely different behavior between M1 and the other models M2-M4 in Fig.75 is discussed further in the following.

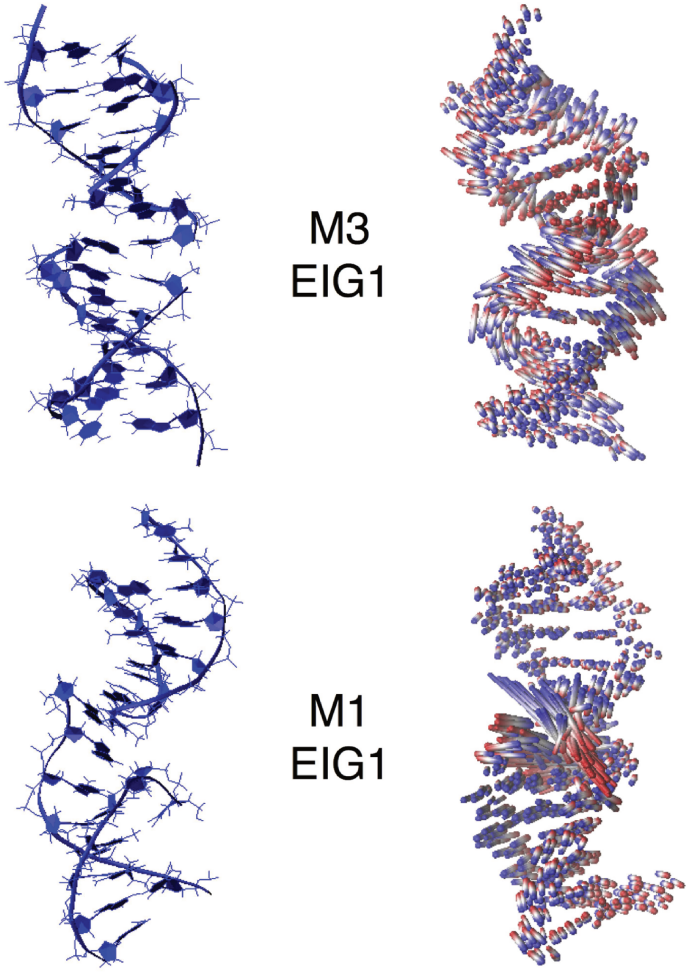


Figure 74: **Configurations associated with first eigenvector of the covariance matrix.** On the left, representative configurations of the DNA fragments close to the DSB, in models M3 (above) and M1 (below). On the right, simultaneous plot of the configurations spanned by the principal motions associated with the first eigenvector, for each model. DNA fragments are aligned with their main axis vertical, the DSB being at the center. The superimposed frames are colored from blue to red, the ordering reflects a virtual motion spanning between the eigenvector extremes. A long stick spanning between the two colors identifies a large motion of the corresponding atom; a shorter stick identifies a local oscillation, of smaller amplitude.

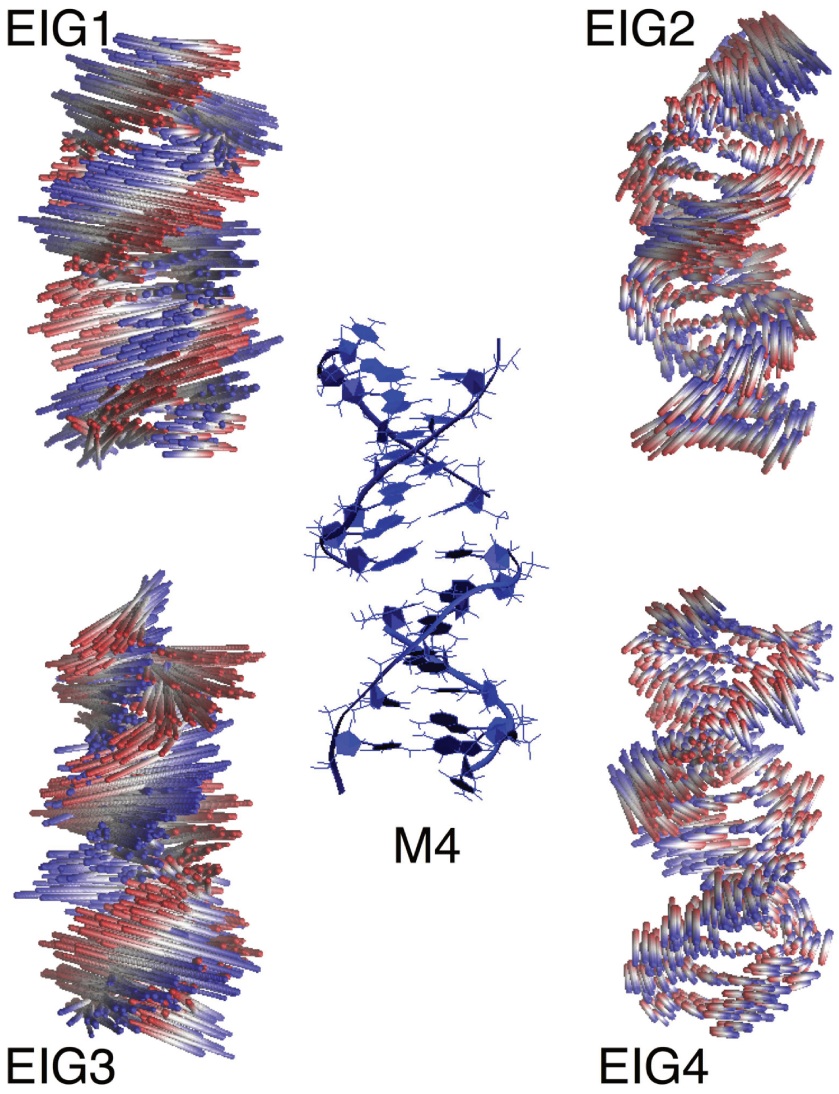


Figure 75: **Configurations associated with eigenvector 1-4 of DSB model 4.** Simultaneous plot of the configurations spanned by the principal motions associated with the eigenvectors 1-4, for the DNA fragment close to the DSB M4 (represented in the central panel). The superimposed frames are colored from blue to red, the ordering reflects a virtual motion spanning between the eigenvector extremes. Also in this case, a long stick spanning between the two colors identifies a large motion of the corresponding atom; a shorter stick identifies a local oscillation, of smaller amplitude. It can be readily appreciated that eigenvectors 1 and 3 correspond to a coordinated, twisting motion of the entire fragment, while eigenvectors 2 and 4 correspond to smaller and less cooperative deformations.

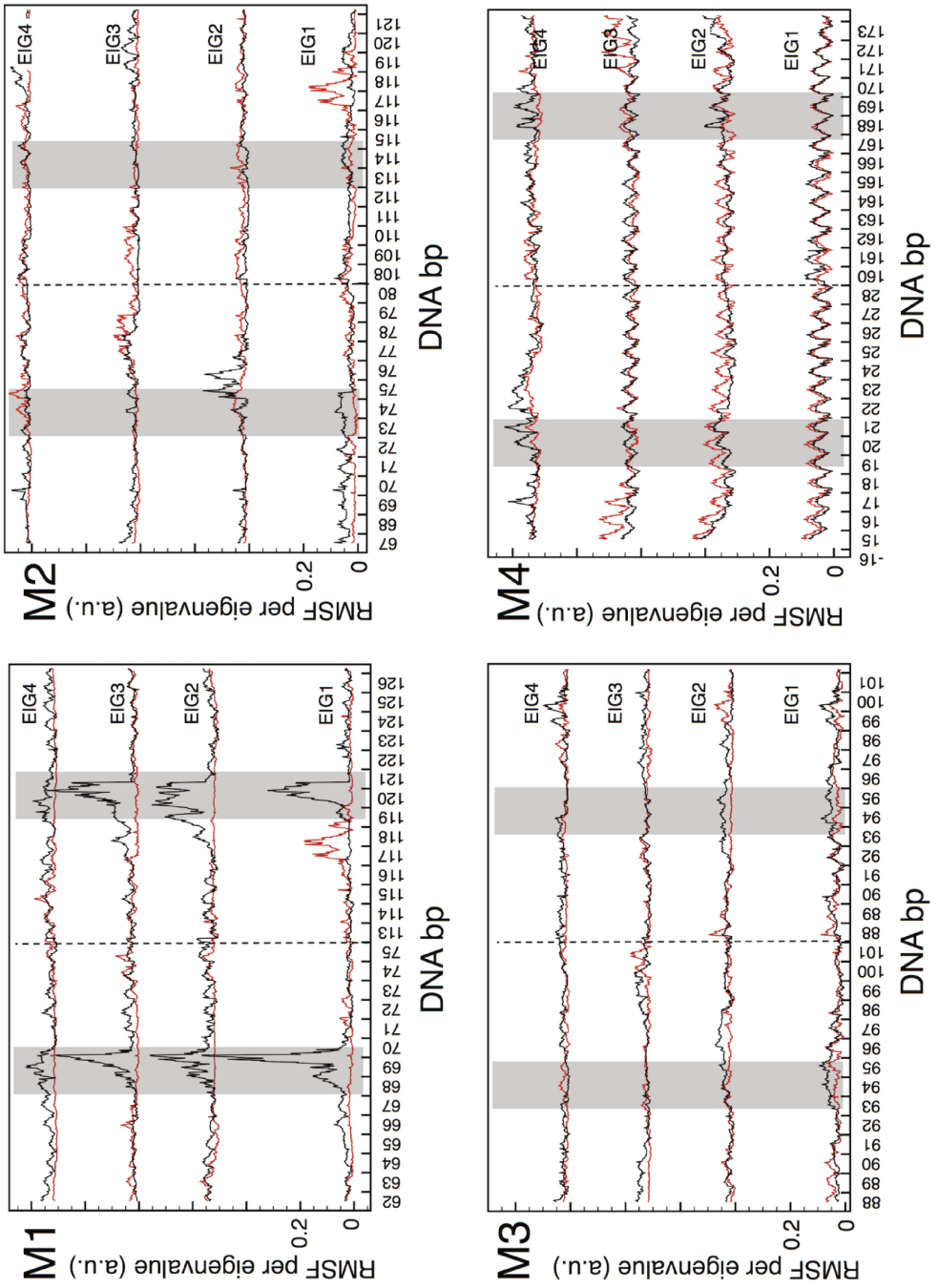


Figure 76: **Eigenvectors of the covariance matrix for DNA at at DSB.** Plot of the RMS fluctuation of the 4 principal eigenvectors, for the DNA fragments including the DSB, of the four models M1-M4. For each eigenvector (respective origins shifted along the ordinate axis) the black line gives the atom-by-atom contribution of the fragment surrounding the DSB, while the red line gives the same quantity for the same fragment intact (from model O). On each abscissa, atoms are grouped by the base number, in two contiguous blocks divided by the dashed line, representing the two parallel strands; the grey shaded regions indicate the position of the DSB (the red-green bases of Fig. 70b) for each model.

In Figure 76 the RMSF for the first 4 eigenvectors of each DSB model are plotted. Each panel in Fig. 76 compares the RMSF for the fragment of 7+7 bp of DNA enclosing the DSB on either side (black lines), with the corresponding RMSF of the same fragment intact (red lines). For the M2-M4 models, it can be clearly seen that the RMSF of the DSB fragments is comparable to that of the same fragment in the reference model O; despite local quantitative variations, also of some importance between the various DNA bases, the black and red traces remain always close to each other, for each eigenvector, within a range of 0.1 in the arbitrary units of the RMSF. Moreover, the regions of the DSB and the base-pairs immediately adjacent (indicated by grey shaded areas) do not seem to display a peculiar or specific behavior, compared to the bp more distant from the DSB locations. Only the 1st and 3rd eigenvectors of M4 are somewhat outstanding compared to all the others, since they display an even distribution of displacements among all the bp.

As it can be verified by looking at the detailed eigenvector plots in the preceding Fig. 75, this coordinated motion correspond to an ample twisting about the main axis, which exists both for the O and M4 model, therefore independently on the presence of the DSB. It may look that the first eigenvector of M4 is more perturbed than the first of M1 (compare Fig. 76); however, the displacements are homogeneous throughout the configuration for M4, whereas for M1 the large motion is concentrated around the 2-3 bp that make up the DSB, which "suck up" the entire eigenvector. Such a difference underscores once more that the displacements defined by the eigenvectors are not true atomic displacements, but relative weights of the total displacement.

This same analysis for the RMSF of the groups of about 16-18 histone residues closer to the DSB in each model, is shown in Fig.77. Also in this case, for the M2-M4 models it is hard to see a qualitative difference between the data for the intact fragments (red lines), and for the fragments with the DSB inserted (black lines). The lysine and arginine residues are overall more mobile than the others, as far as the 4 principal eigenvectors are concerned, describing a dynamic interaction with the DNA. However, with minor variations, this behavior is the same also in the absence of the DSB, therefore it reflects the usual affinity of such residues for the DNA bases. The M1 model, instead, is definitely different, as it was the case for the DNA analysis in Fig. 76 above, and it will be treated later in this Section.

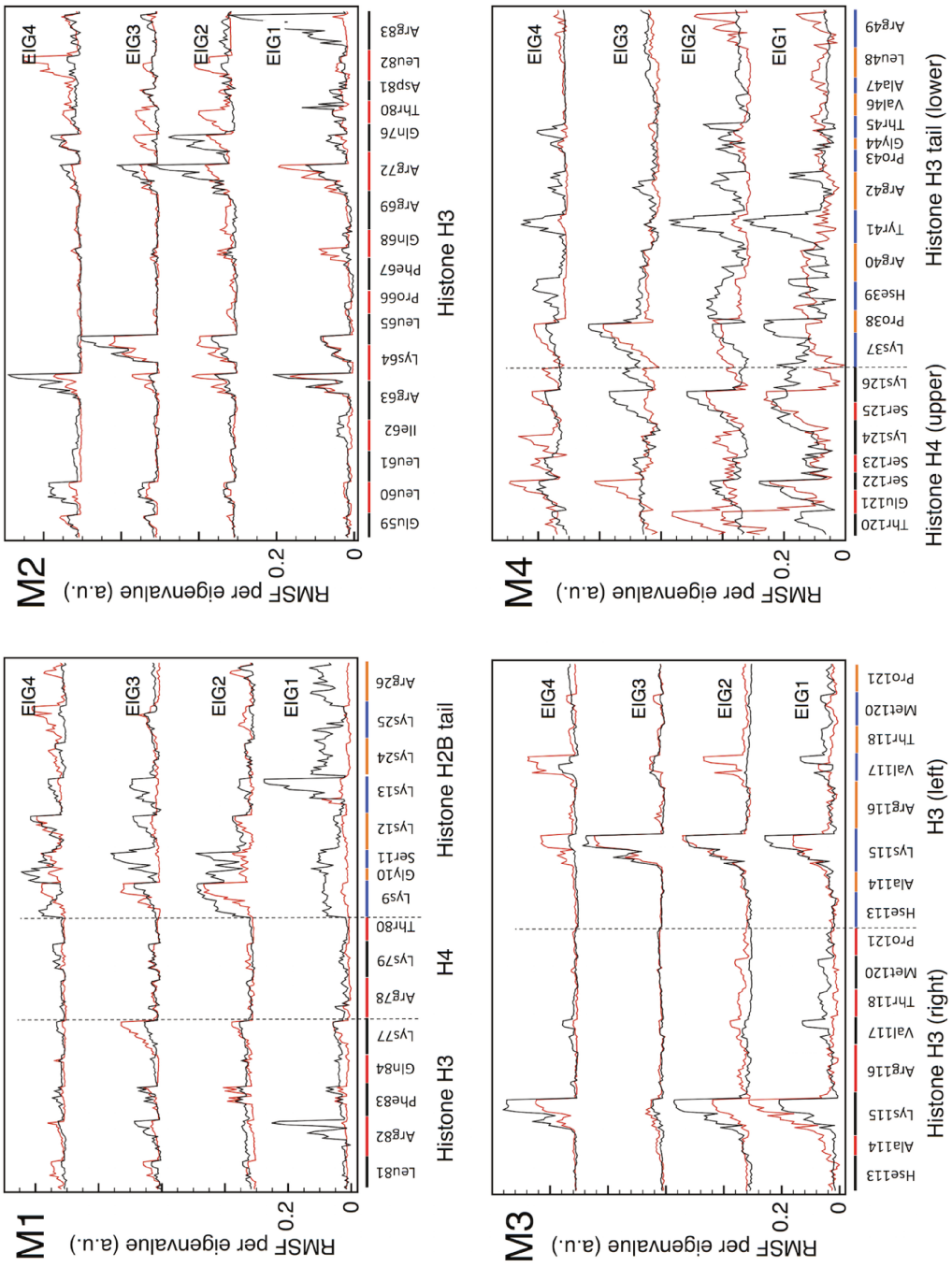


Figure 77: **Eigenvectors of the covariance matrix for histones at at DSB.** Plot of the RMS fluctuation of the 4 principal eigenvectors, for the histone residues closest to the DSB in each of the four models M1-M4. For each eigenvector (respective origins shifted along the ordinate axis) the black line gives the atom-by-atom contribution of the fragment surrounding the DSB, while the red line gives the same quantity for the same fragment intact (from model O). On each abscissa, atoms are grouped by residues, with a spacing (also indicated by colored bars) proportional to the size (i.e., number of atoms) of each residue.

The Schlitter entropy formula [135] can be used to estimate an upper limit for the contribution to the free energy from the excess entropy due the presence of the DSB, as:

$$T\Delta S_{DSB} = T\Big(\langle S_{MX}\rangle - \langle S_O\rangle\Big) \tag{7.1}$$

with $MX = M1, ...M4$, and $\langle ...\rangle$ indicating the time average of the Schlitter entropy for each molecular fragment:

$$S = \frac{1}{2}k_B \ln\left\{\det\left[\mathbf{I} + \frac{k_B T e^2}{\hbar^2}\mathbf{MC}\right]\right\} \tag{7.2}$$

with **C** the covariance matrix of the atomic displacements, **I** the identity matrix and **M** the mass matrix, having respectively 1 and the atom masses on their diagonals, and 0 elsewhere. Table 8 reports the values for each DSB model, divided into DNA and histone contribution.

The absolute DNA entropy S_O from Eq 7.2 fluctuates about 18±0.5 kcal/mol/K for each bp, very homogeneously all along the most part of nucleosome, but increasing to 20 kcal/mol/K in the few terminal bps attaching to the straight segments. If the values of excess entropy of DNA are distributed to the 4 bases (green and red in Fig 70b) comprising the DSB, these correspond to an excess of 35 to 60% for the M2-M4 models, the excess per base being larger in the M4, in agreement with the somewhat larger mobility demonstrated in Fig. 76. On the other hand, the excess entropy for the histone residues selected for this analysis remains relatively small, for the three models M2-M4. Despite some difference in the total masses of the groups selected, even when expressed per unit mass instead of per-moles, the absolute entropy of the histones remains comparable, between the model O and the models including the DSB. This is a further confirmation of the relatively minor role played by histone dynamics in the M2-M4 models.

Table 8: **Excess entropy of DSB fragments.** Upper limit of the excess entropy contribution $T\Delta S$ to the free energy at T=310 K, estimated from the Schlitter formula, Eq 7.2, for each molecular fragment in the different DSB models.

DSB configuration	DNA (kcal/mol)	histones (kcal/mol)
M1	89.7	29.8
M2	25.6	6.9
M3	38.1	2.7
M4	47.9	5.3

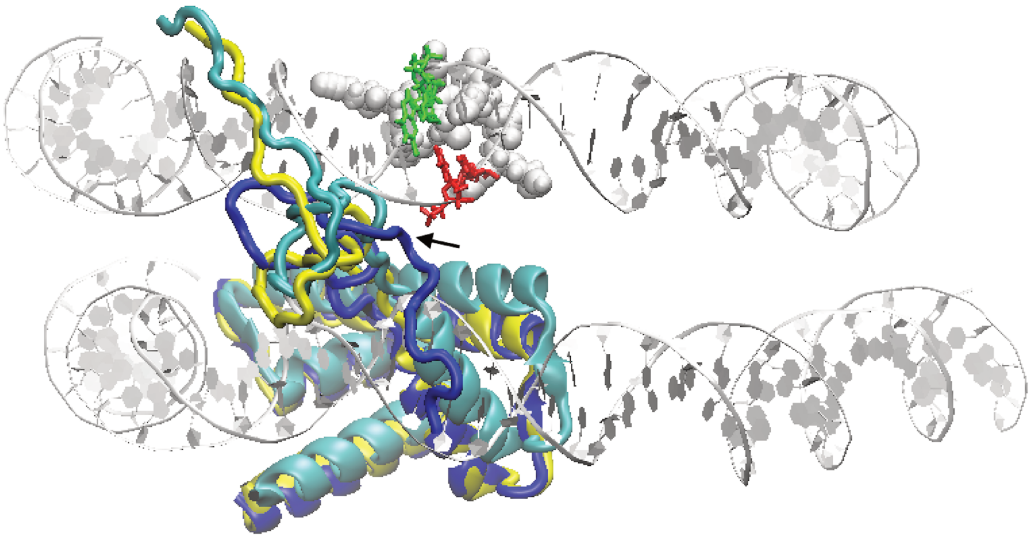


Figure 78: **Different conformations of histone H2B tail at a DSB.** Representation of the time-averaged H2B histone configurations: in the reference O model (yellow ribbons), in the M1 model at T=310 K (cyan) and in the M1 model at T=350 K (blue); the tail of the histone can be seen taking different orientations in the former two, w/r to the last simulation. Two portions of the upper and lower turn of DNA about the histone core are represented in silver ribbons; the spheres on the back represent the H3/H4 residues closer to the DSB; the DSB bases are colored in red-green according to the scheme in Fig. 70(b).

We now turn to describing the behavior of the DSB in the M1 model. Contrary to our expectations, this location in which the DSB is constrained between the histone core and the mobile H2B tail, and close to a DNA-protein contact, was the one to display the most interesting dynamics. The most evident change in the immediate environment of the DSB is the modification of the H2B tail, which can fold into very different interacting positions, starting from the outward extended conformation of the experimental crystallographic structure.

This behavior can be appreciated in Figure 78 above, where the arrangement of the H2B tail is represented for three configurations, averaged over the respective MD trajectories: the reference O model (yellow), the M1 model at T=310 K (cyan), and the M1 model at T=350 K (blue). The low-temperature average configuration of the H2B tail resembles well that of the O model, with the terminal wrapping the minor groove of the DNA strand on the left of the DSB (in the figure); the high-temperature average configuration, instead, has the H2B tail flipped down by about 180 degrees (occurring very early in the trajectory, and irreversible over the whole 1.8 μs), with the fold of Lys24-25 and Arg26 keeping close contact with the DSB (see the black arrow). That such a configuration may be dynamically sampled over ~1 μs time by only a 40 K temperature difference, means that the corresponding energy barrier (chemical plus deformation) must be relatively small.

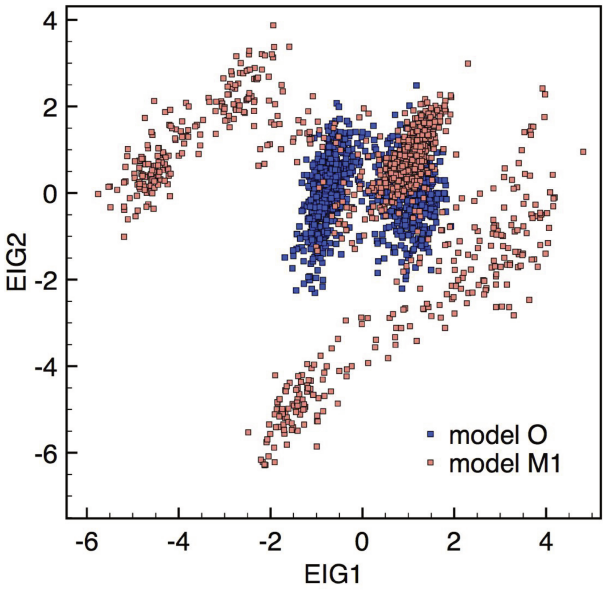


Figure 79: **Projection of the MD trajectory on the two principal eigenvectors.** The plot shows the projection of the ~1μs MD trajectory for the DNA fragment comprising bp 62-113 to 75-126, for the pristine nucleosome model O, and for the model M1 including a DSB at this site.

In this M1 model, the DSB is constantly enclosed between the two β-sheets of H3 and H4, which fluctuate about their equilibrium structure and interact with one side of the DSB, while the H2B tail experiences strong oscillations, coupling with the cut bases of the opposite DSB side. The time evolution of the four bases comprising the DSB (green-red colored in Fig. 78) gives a qualitative appraisal of this strong interaction. Notably, the interacting portions of both the two β-sheets, and the H2B tail, include more than 60% of lysine and arginine residues, as expected given the strong electrostatic affinity of such amino acids for DNA (notably for G and T, [152]). The DNA ends at the DSB are clearly perturbed by such interactions, and it can no longer be said that the two broken backbones preserve a geometrical continuity, as it was instead observed for the M2-M4 models for the entire duration of the respective MD trajectories.

By looking at the eigenvalue RMSF for the M1 model in Fig. 76, it can be seen that in this case the group of DNA bases adjacent to the DSB take up the majority of the weight, indicative of their participation in the ample fluctuations of the open DSB ends. It is important to note that the types of motion described by the first few eigenvectors are definitely different between the DNA with, or without the DSB, as demonstrated by the projection of the entire MD trajectory on a virtual plane whose (x, y) axes are the first two eigenvectors (Figure 79); each point represents a pair of values $(\mathbf{v}_1, \mathbf{v}_2)$ from MD frames spaced by 10 ps, the blue and red dots representing the trajectory of model O and M1, respectively. Apart from the small region [0:2,0:2], most likely due to the less mobile base-pairs in the fragment, there is practically no superposition between the essential subspace of the

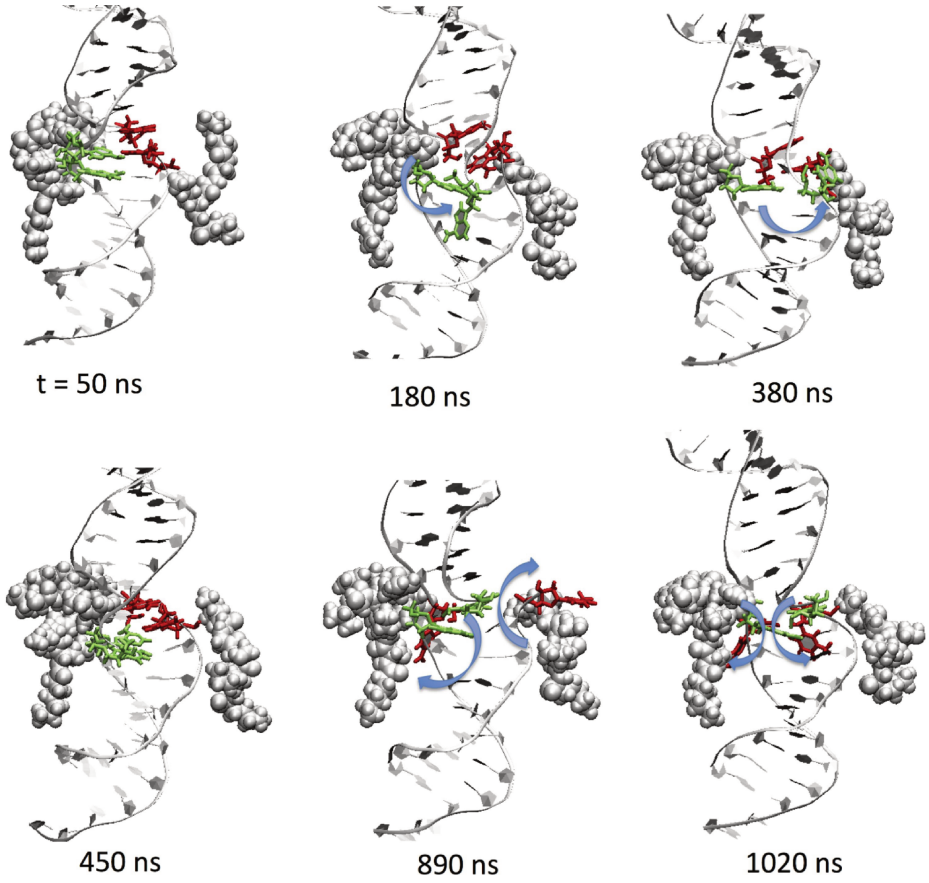


Figure 80: **Time-sequence of the evolution of the DNA-histone contacts** for the DSB at the M1 position. The groups in grey VdW-spheres are the Leu82-Arg83-Phe84-Gln85 of H3, Lys77-Arg78-Lys79-Thr80 of H4 (left side); and Lys9-Gly10-Ser11-Lys12-Lys13, Lys24-Lys25-Arg26 of H2B (right side). DNA bp around the DSB are colored red-green according to the scheme of Fig.70(b).

two DNA configurations, meaning that such large-scale motions are almost entirely different between the two simulations.

The motion corresponding to $\mathbf{v}_1$ in M1 is essentially carried by the few base-pairs making up the DSB (see again Fig. 74): this type of motion (eigenvector) does not exist in the intact fragment. The component analysis for histone eigenvalues in Fig. 76 confirms that, similarly to all models, the principal motions are evenly shared by all atoms in the absence of the DSB, whereas the high-energy dynamics becomes fully localized around the structural defect when the DSB is present (Figure 80), carried especially by the lysine and arginine residues. The values of excess entropies from Table 8 provide a further confirmation of the peculiar large-scale dynamics of this DSB configuration. Because of these indications, we extended the MD trajectory of the M1 model up to 1.8 μs, but never observed a true mechanical destabilization of the DNA structure: the two DSB ends remained firmly in place, even if the two terminal base-pairs on each end fluctuate quite wildly (see again Fig. 74, and the motion indicated by blue arrows in Fig. 80), while promoting a strong interaction with the protein surfaces.

7.3 STEERED-MD AND UMBRELLA SAMPLING

As shown in the preceding Section, spontaneous dissociation of one or both DSB ends of a broken DNA from the nucleosome remains a difficult event, never observed in our simulations. DSB opening, and DNA detachment from the nucleosome are likely governed by a free energy barrier of adhesion, which even such a critical defect as a fully-cut DNA could not easily overcome simply by thermal fluctuations. A way to estimate the free-energy barrier in such a large and complex molecular system is to resort to controlled-force pulling, in order to impose the detachment, and then to use the intermediate structures along the reaction coordinate as starting points for the "umbrella" sampling of the potential of mean force (as described in Section 4.3.4). From the latter analysis, the free energy barrier(s) along the chosen reaction coordinate can be extracted.

Steered molecular dynamics (SMD) was performed on the fragments with the constant-force pull code available in GROMACS, only on the M1 model. In this case, we enlarged the water box to 18 nm in the x-direction, to allow possible outward extension of the broken DNA end, resulting in a system of 107,000 water molecules. Since the objective was to promote the detachment of one of the broken DSB ends from the nucleosome core, we applied a constant force parallel to the direction x and perpendicular to the superhelical axis (see Figure 81(a)), by means of a harmonic-spring fictitious potential attached to the $C4'$ and P atoms of the last two base pairs at one DSB end. After some tests, the spring constants were set at 100 and 75 kJ mol^{-1} nm^{-2}, respectively for the two DNA strand-ends farther and closer to the nucleosome surface. To provide a reaction force keeping the system in place, all the atoms of the H3 histone opposite to the DSB were retained by soft harmonic restraints, with a spring constant of 250 kJ mol^{-1} nm^{-2}; pulling speeds of 1 to 5 m/s were used for SMD simulations; forces and displacements were recorded at intervals of 5-10 time steps. Umbrella sampling was performed by extracting 100 configurations spaced by 50 ps during the first 5 ns of the force pulling simulation; the force bias was progressively reduced from 100 down to 10 kJ mol^{-1} nm^{-2}, to extrapolate to the zero-bias limit of the free-energy profile; the weighted-histogram analysis was used to interpolate and connect the data from the sequence of discrete configurations.

As the reaction coordinate ζ, we took the separation distance between the moving DSB end and the histone core surface (see again Fig. 81(a)). This was measured by taking the center of the DNA axis, at the average position of the $C4'$ and P atoms of the last two base-pairs, and projecting it on the closest histone surface atom, along the line perpendicular to the superhelical axis. Figure 81(a) also shows the variation of ζ as a function of simulation time, at constant pulling force. It can be seen that the DNA broken end detaches from the histone surface in large steps (red segments), during which the internal energy builds up until some barrier is overcome; the final stage, indicated by the blue segment, is the complete detachment of the DSB

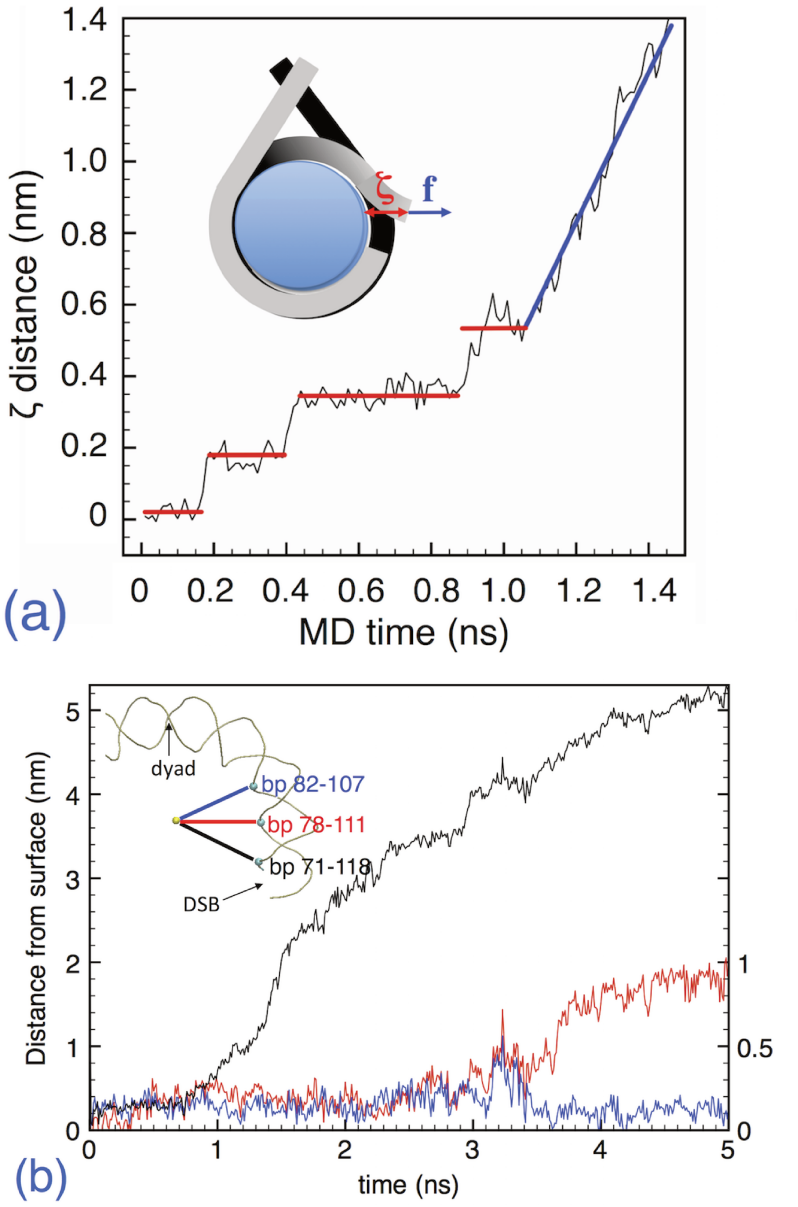


Figure 81: **Force-induced detachment of a DSB end from nucleosome. (a)** Plot of the reaction coordinate ζ as a function of the MD simulation time, for a pulling simulation at constant-force resulting in the slowest average velocity of about 1 m/s. The inset shows the definition of the ζ distance of the broken DNA end (red arrow), and the direction of the applied force vector (blue arrow), parallel to the x axis and perpendicular to the superhelical axis. **(b)** Detachment of DNA portions from the histone core surface, from 0 to 5 ns of MD simulation at constant force and 310K. The three traces (black=left ordinates; blue/red=right ordinates) represent the distance to the surface of the three P atoms indicated in the scheme on the upper-left corner, in which a quarter of turn of the DNA comprised between the dyad and the DSB is sketched. The yellow sphere is the geometric center of the nucleosome.

end after t=1.1 ns, in which the free end is simply drifting at the constant speed of about 2 m/s (later on dropping to 1 m/s).

During the final stage of the pulling simulation, the DNA is forcefully unwrapped from the histone core, as it can be seen in Fig. 81(b). Here it is shown the distance from the core surface of three P atoms facing the histones, belonging to the bp 71-118 (contact site close to the DSB), 78-111 (middle site) and 82-107 (next contact site). The first contact site is detached in the interval t=1.-1.5 ns, as indicated by the black trace that follows the distance from the surface of of the P71 backbone phosphor. Then, under the continued pulling of the DSB end, also the P111 comes off, at t >3 ns (red trace); however, it may be noticed that this event is "cooperative", the P82 (blue trace) following the instantaneous opening of P111 at t=3.-3.4 ns, and then falling back into position, after which P111 is definitely "peeled off" the histone surface.

7.3.1 *Free energy to detach broken DNA ends*

From this force-pulling simulation we can calculate the free energy profile of the barriers, which characterize the binding of the DNA end to the histone core surface. The potential of mean force (PMF, [69]) is a method to extract the free energy difference ΔG from a sequence of configurations, biased along a reaction coordinate that brings the system from a state a to a state b. In our case, the reaction coordinate is just the distance ζ defined above; the states a, b respectively represent the initial configuration at ζ=0, with the DSB end still attached to the histone surface, and the final configuration with the end detached, at ζ ~5 nm. The "umbrella sampling" technique [159] is used to obtain the PMF at discrete values of ζ, and the discrete values of $G(\zeta)$ between a and b are connected by the weighted-histogram analysis (WHA) [67, 84]. We extracted 100 configurations from the force-pulling simulation, spaced by 50 ps in the first 5 ns of the trajectory (corresponding to about 0.5 Å spacing along the reaction coordinate ζ=0 to ζ ~5 nm); each configuration was equilibrated for 2 ns at 310 K under constant-{NVT}, while biased with a harmonic "umbrella" potential of variable strength, progressively reduced to zero to obtain the unbiased limit. The force probability distribution of the fluctuating DSB free-end at each value of ζ was reconstructed by WHA, and the free energy profile thereby extracted is shown in Figure 82.

Despite the noisy profile, a few features can be identified. The red circle defines the first barriers to the detachment of the DSB ends, corresponding to the red steps in Fig. 81(a); such barriers are quite small (<1 kcal/mol), and strongly depend on the choice of the point of application of the pulling force. The blue circle identifies the free-energy barrier for the detachment of the first contact at P71, about ΔG=1.8±0.2 kcal/mol or 3 $k_B T$; this does not represent a very large value, and should correspond to a ~5% Boltzmann probability of spontaneous detachment at T=310 K. It is worth noticing that this value for the detachment barrier fits very well with the experimental estimates of nucleosome unfolding energy, which obtain a value of about 27 kcal/mol [100, 170]: this corresponds to the detachment of

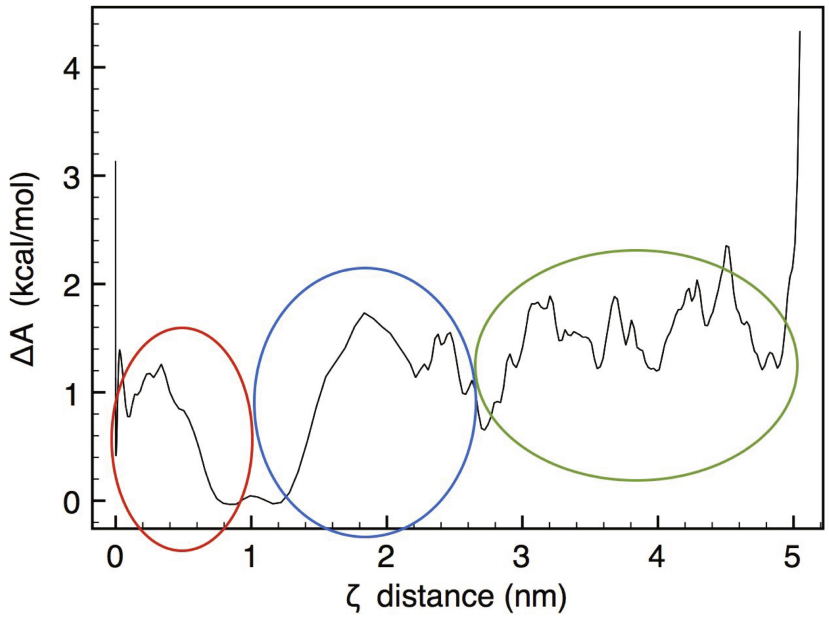


Figure 82: **Free energy for DSB detachment.** Zero-force extrapolated free-energy profile as a function of the ζ reaction coordinate, during the force-pulling simulation at T=310 K. The red, blue and green regions refer to the sequence of energy barriers for the DNA-histone detachment events, as described in the text.

all the 14 contact sites, from which it can be estimated an average energy of 1.9 kcal/mol per contact. The green circle roughly identifies the cooperative events leading to the detachment of P111, between 2 and 4 ns, with a sequence of ΔG, again, not larger than 2-3 $k_B T$. Further detachment events were not observed, with the above values of pulling force; in particular, P82 remained in place for >500 ns, even at larger deformations of the DSB free end, because of the H3 histone tail acting as a sort of brace that maintains the DNA firmly in place about that position. Much larger forces, or cooperative events of histone tail fluctuation, likely involving other nuclear proteins, seem to be necessary to pull the free DSB end further beyond the limits observed in the present simulations.

7.4 MOLECULAR STRESS CALCULATION

In the last part of our study on the nucleosome, we turn our attention to the internal relaxation dynamics of the nucleosome including a broken DNA. To demonstrate what it is meant by "internal relaxation", we take two configurations along the final trajectory of the force pulling simulation of the M1 model described in Fig. 81, C180 and C290, respectively extracted at times t=1.8 and 2.9 ns, well beyond the detachment stage that ends at 1.1 ns in the figure. Each of these two configurations is used as initial structure for an MD simulation, and is then equilibrated and relaxed at 310 K and constant-{NVT}, without any external forces applied. The results of these two MD simulations are displayed in Figure 83: starting from the two

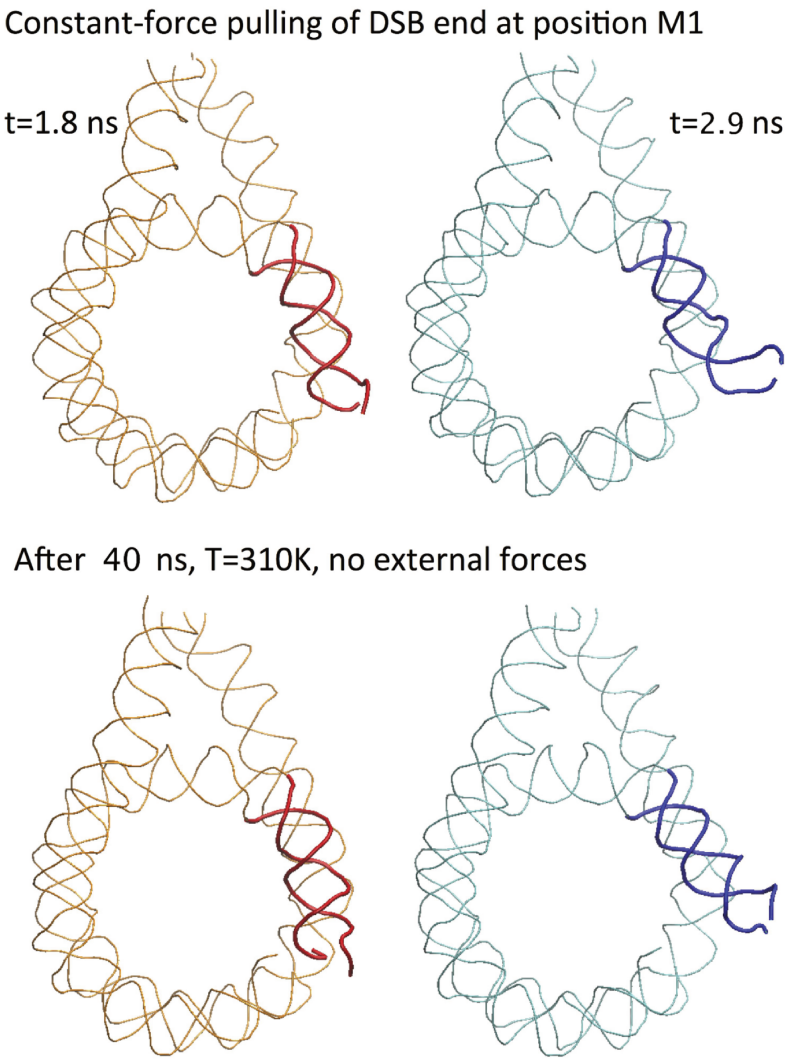


Figure 83: **Stress relaxation in a detached DSB**. Two configurations extracted from the force-pulling MD simulation of DSB at position M1, at t=1.8 ns (C180, left, red ribbons) and t=2.9 ns (C290, right, blue ribbons). The pull force was applied only at the C′-P atoms of the 2 last bp on the upper end of the DSB. The terminal portions of the pulled DSB end are highlighted as a thicker tube, for clarity. Row above: the two configurations at the start of the relaxation. Row below: the two configurations after 40 ns of MD equilibration/relaxation at T=310 K without any external forces applied.

different initial conditions, after 40 ns the C180 tends to fold back into the initial M1 configuration, while C290 straightens out and increases its distance from the histone core. Notably, the C180 remains in a slightly open state, because of the free energy barrier to detachment that now has to be overcome in reverse. However, the important observation in both cases is that the folding back, or the straightening out, are driven entirely by the competition between the residual attraction between DNA and proteins (a "chemical" force), and the relaxation of internal constraints (mainly bending and torsion, therefore an "elastic" force).

The role of internal forces can be clearly understood by looking at the distribution of **mechanical stress**, which is a measure of the elastic energy accumulated by the bending and torsion of DNA while wrapping around the histones, and that is ready to be released if the structural constraints are softened, as it could be the case of a DSB cutting the DNA sequence.

The "molecular" definition of stress proposed by Cauchy around 1828 [27]:

$$\sigma(\vec{r}) = \frac{1}{V}\sum_i \left(\frac{\vec{p}_i \otimes \vec{p}_i}{m} + \frac{1}{2}\sum_{j\neq i} \vec{f}_{ij} \otimes \vec{r}_{ij} \right) \tag{7.3}$$

was based on a continuum-mechanics representation of the forces $\vec{f}_{ij}$ between pairs of point-like "molecules" of mass m at distance $\vec{r}_{ij}$: by considering the infinitesimal volume around a point $\vec{r}$, the stress would be the result of the eventual imbalance between the "molecular bonds" crossing in and out its bounding surface. However, this definition presumed a strictly homogeneous system of total volume V, subject to small deformations so that linear elasticity theory applies; the first term relative to molecular momenta $\vec{p}_i$ was absent in Cauchy's original definition, and was introduced much later.

Extracting an observable equivalent to Cauchy's proper mechanical stress from a molecular simulation is a subject that has attracted great interest, as well as sharp controversy (see, e.g., [1, 29, 33, 79, 96, 150]). We will not step in the complexity and subtlety of the arguments, because this would represent a too large detour from the objectives of this work. In MD simulations it is customary to use the so-called "virial" definition of the stress, stemming from Cauchy's expression above but projected down to an atomic volume; the shortcomings of this poorly justified, empirical approach have been repeatedly underscored.

Notably, recent developments led to alternative geometric derivations of the microscopic stress [154, 158], based on the invariance of the free energy with respect to surface deformations [45, 109], instead of the classical formulation based on invariance of momentum. The so-called *covariant central-force decomposition* scheme (CCFD, [157, 158]) for the intra- and intermolecular forces, already incorporated in the GROMACS code by M. Arroyo's group in Barcelona (`http://www.lacan.upc.edu/LocalStressFromMD`), ensures conservation of both linear and angular momentum under a generic stress-induced volume transformation. The method is implemented in a special-purpose patch to GROMACS 4.6, which reads (all, or part of) a MD trajectory for the selected subset of atoms for which stress is to be computed, and performs the entire analysis. Since the GROMACS-LS patch [157, 158] constrains the code to run in serial rather than in parallel, care must be taken to define properly the subset of interest in order to avoid prohibitive computing times.

We prepared a set of scripts and subroutines to extract the principal components of the stress, to compute scalar projections, to compare stress fields from different simulations, and to reformat the outputs

in the portable Gaussian-cube format for visualization. Generally, the stress field generated by CCFD was averaged over trajectory segments of 1 ns, with 100 frames spaced by 10 ps. This choice is a compromise between obtaining significant statistics while reducing the noise: in fact, averaging over a longer time window would progressively smear out the differences, while averaging with more frames separated by shorter interval would progressively increase the noise. Comparison between stress fields from different MD runs poses an extra care, since the structures need to share exactly the same box size and center, to avoid numerical artefacts from the cancellation between large positive and negative values. According to the CCFD scheme, stress fields are calculated by GROMACS-LS on a continuous grid superposed on the molecular structure; however, stress components and individual force contributions (pair, angle, dihedral, etc.) can also be projected back on the atom sites, by previously defining a conventional (but non unique) atomic volume.

7.4.1 *Internal stress relaxation and DSB structure*

The mechanical stress $\sigma(\vec{r})$ (a 3×3 tensor defined at any point $\vec{r}$ in space) is a meaningful way of representing the distribution of internal forces with respect to a given local direction vector. Once a DSB breaks the DNA backbone around the nucleosome, internal forces are going to be relaxed, and compete with the chemical (Van der Waals, electrostatic) forces from the interaction with the histone proteins. Looking back at Fig. 83 for model M1, such a competition is very evident upon comparing the bottom configurations: in C185 the chemical forces overwhelm the internal stress, whereas in C290 the opposite holds, and the DNA ends up straightened out from the DSB site.

A tensor, such as the stress, can be meaningfully projected onto any direction vector, the choice of a particular projection being just a matter of convenience. In the present case, the "bent tube" structure of nucleosomal DNA makes it interesting to consider the stress projected onto its "tubular" surface (see the scheme in Fig. 70(c)). Stress projections are scalar quantities, which can be more easily visualized compared to 2-D or 3-D density fields.

An intuitive way of looking at the mechanical stress as a "projected force" is through the surface traction vector:

$$\vec{T}(\vec{r}) = \sigma(\vec{r}) \otimes \hat{n} \tag{7.4}$$

The symbol $'\otimes'$ indicates the tensor product between the stress and the vector $\hat{n}$, in practice the matrix product between the 3×3 matrix of the stress at each point $\vec{r}$, and the 3-component vector locally perpendicular to the surface at the point $\vec{r}$ (see the local reference frame $\{\hat{n}, \hat{t}, \hat{b}\}$ in Fig 70(c)). The traction vector $\vec{T}(\vec{r})$ (exactly the original Cauchy's definition of stress vector as "flux of momentum across an infinitesimal surface element") contains a great deal of information on the state of internal tension, compression, and torsion, of a complex structure like the DNA in the nucleosome.

Notably, the portion of DNA wrapped around the histone core in the nucleosome is forced to bend into nearly two full circles of diameter about 8 nm, a size much shorter than the persistence length of free DNA, $\xi_p \simeq 50$ nm. Therefore, the DNA "tube" is here constrained in a geometry from which it should rather escape into a more straight structure, whenever possible, under the relaxation of internal forces. The state of tension and compression of a bent tube can be described by this particular projection of the traction vector:

$$t(\vec{r}) = \vec{T}(\vec{r}) \cdot \hat{\tau} = (\sigma(\vec{r}) \otimes \hat{n}) \cdot \hat{\tau} \tag{7.5}$$

in this case along the unit vector $\boldsymbol{\tau}$ locally tangent to the continuous line sweeping the center of the tube.

Notably, a bent tube would experience a stretching force (a tensile, negative $t(\vec{r})$) in the half that lies outside the centerline with respect to the center of curvature, and a compressive force (a positive $t(\vec{r})$) in the half lying inside the centerline, as shown in blue/orange in Fig. 70(c). The internal force should be zero along the centerline itself, because of this called the "neutral axis" (also the helical axis of the circularly-bent DNA).

We computed the line tension $t(s)$ all along the curved DNA pathlength s spanning the length of the helical axis, by averaging over slices of width 0.5 nm (see for example the white slice in Fig. 70(c)), and by integrating separately over the inner and outer regions (orange and blue in the Figure). Each slice averages all the points $\vec{r}$ included in the white volume section, centered at the midpoint between the two P atoms of each base-pair; therefore adjacent slices have some geometric overlap, to provide a smoother profile of the signal. Note that in a perfectly smooth, homogeneous tube, one should see just two constant values of positive and negative tension, respectively in the blue and orange volumes. However, the DNA is not simply a smooth tube, but it has a complex geometry in which minor and major grooves alternate, and it contacts the histone surface in about 14, evenly spaced sites. At these points, there is an excess or a defect of tension/compression, as well as some amount of under/over twisting of the already twisted tube.

The twist stress is that part of the internal forces involved in the torsion about the central (neutral) axis of the tube. The DNA double helix is naturally twisted already in its normal B-configuration; however, when it is bent in the nucleosome, the twist is necessarily modified with respect to the normal configuration. The twist component is obtained as well from the traction vector, as:

$$w(\vec{r}) = \vec{T}(\vec{r}) \cdot (\hat{\tau} \times \vec{r}) = (\sigma(\vec{r}) \cdot \hat{n}) \cdot (\hat{\tau} \times \vec{r}) \tag{7.6}$$

where $\vec{r}$ is the position vector computed from the neutral axis, and parallel to the local surface normal $\hat{n}$ (see again Fig. 70c). The vector product between $\hat{\tau}$ and $\vec{r}$ defines a third vector parallel to the local tangent, threading like a spiral screw about the DNA tube; positive and negative values of $w(\vec{r})$ indicate a rotational force (a *torque*) tending to over- or under-twist the DNA about its helical axis.

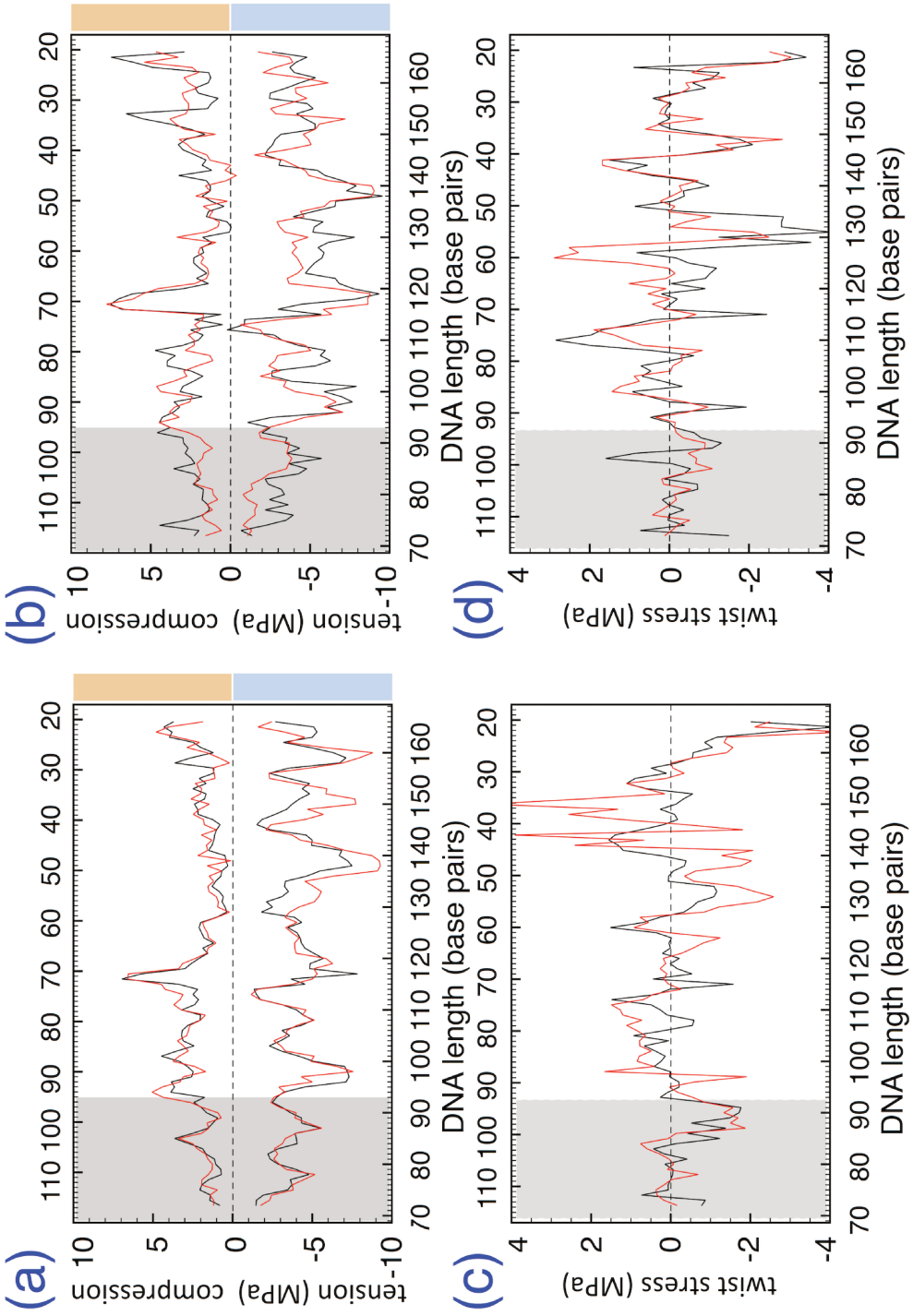


Figure 84: **Stress profiles along DSB terminal ends for the M1 model. (a)** Plot of the tension/compression force $t(s)$ along part of the DNA model M1 configuration C185, at the beginning of the relaxation (black line) and after 40 ns (red line). Numbers along the upper and lower ordinates indicate the DNA bases in the two strands; the DSB is at the extreme left, bp 68-120; the grey rectangle is the main stress relaxation region, from the dyad position at bp A94···T94; color bars on the right refer to the two regions in Fig 70c. **(b)** Same as (a), for the M1 configuration C290. **(c)** Plot of the twist force $w(s)$ along the DNA model M1 configuration C185. **(d)** Same as (c), for the M1 configuration C290.

In Figure 84(a) and 84(b) we show for both configurations the tension profile $t(s)$ along the helical axis of the DNA fragment right after the DSB (bp T68$\cdots$A121). The portion mostly affected by the pulling under force and subsequently relaxed is comprised between the DSB and ~bp T84$\cdots$A105; we neglect the first few bp immediately next to the DSB, too disordered for such a calculation. Two sets of data are shown in each panel, at the beginning of the relaxation (black lines), and after 40 ns (red lines); stress values are averaged over 100 frames with 10 ps spacing, in either case. In general, the terminal part of the DNA next to the DSB (indicated by a grey-shaded area in the panels) tends to lower values of both line tension and compression, for both configurations, compared to the rest of DNA beyond the dyad (bp A94$\cdots$T94), indicative of the stress release at the free ends. The extra tension/compression from the DNA-histone contact points can be clearly observed in the alternating minima and maxima along the compression and tension sides of the DNA tube.

Despite some noise in the data, it can be appreciated that for the C185 configuration (Fig. 84(a)) the red lines are at the same values than the black ones: this is a signature of the chemical residual attraction winning over the internal stress, thus tending to fold back the DSB open end into place. On the other hand the C290 in Fig. 84(b), starting from almost twice-higher stress values in the grey area compared to the C185, has red lines approaching a state of nearly zero stress after the relaxation time; also several sites beyond the dyad (outside the grey region) display sizeable variations of tension and compression. This suggests the release of internal stress as being responsible for straightening out the DSB end, into a mechanically less-constrained structure.

The extra twist stress (positive or negative) also contributes to the internal forces that are going to be relaxed, when the DSB cuts open the DNA, albeit to a much lesser extent, given the smaller absolute values of w compared to t. In Fig. 84(c) and 84(d) the $w(s)$ stress profiles are shown, under the same conditions of the two panels above for the line-tension/compression. It can be noticed that, also for the twist stress, generally smaller values ($\leqslant$1 MPa in modulus) are observed in the DSB tail. However, the large numerical noise does not allow in this case to draw a more firm conclusion, concerning the (likely minor) role of twist stress in the chemical vs. mechanical force competition between the two configurations.

7.5 FINAL SUMMARY

In this Chapter, we studied by very-large-scale MD simulations the evolution under external force and temperature of DSBs in nucleosomal DNA. We collected and analyzed a large amount of raw data (more than 1.5 TBytes, and more than 5 million CPU hours on two large supercomputers), by running microsecond-long trajectories for five different, all-atom models of the experimental 1kx5 nucleosome structure [41]. The basic model is made up of the canonical 8 his-

tones, plus a 187-bp DNA comprising the 147 bp wrapped around the histone core and 20-bp terminations on each end, and embedded in large boxes of about 80-110,000 water molecules with Na^+ and Cl^- ions at 0.15 M physiological concentration. The pristine nucleosome configuration (model O) was modified, by inserting a DSB at four different positions in the DNA (models M1-M4), and the stability of the resulting structures was compared with model O nucleosome.

A general observation from the μs-long trajectories, is that damaged DNA remains well attached to the nucleosome body, without major qualitative differences compared to the intact DNA. Only the model M1, in which the DSB is tightly sandwiched between the histone H3 and the tail of histone H2B, displayed a dynamics substantially different from the corresponding region in model O, due to the increased interaction of the broken DSB ends with close-by histone residues; however, also this DSB configuration was stable over the entire observation time scale, which in this case was extended to 1.8 μs.

In order to identify the free-energy barriers which maintain the broken DNA attached to the histone core, we carried out steered MD with a pulling force to "peel off" the free DSB end from the nucleosome; relatively small free-energy barriers of the order of 3 k_BT were identified, which could allow spontaneous DSB end detachment at physiological temperatures, likely over longer time scales, of hundreds of microseconds to milliseconds. At the same time, histone tails represent a major steric obstacle for unwrapping of larger DNA sections. Spontaneous unwrapping of DNA from the nucleosome core has been studied experimentally [89, 122, 160], because of its relevance in gene regulation and DNA transcription; notably, such experiments were carried out on isolated nucleosomes, with a length of DNA just matching, or barely longer than needed to wrap the histone core (147 to ~180 bp). In such conditions, spontaneous detachment of the ends was indeed observed over the timescale of hundreds of milliseconds; simulations by coarse-grained MD methods roughly confirm such trends [48, 78, 163], despite being strongly dependent on the empirical parametrization of each different force model. To such experiments it may be objected that the nucleosome constrained in the chromatin could have a rather different mechanics, with respect to isolated nucleosome particles.

Indeed, our molecular-stress calculations demonstrate that the circularly bent DNA has a strong internal driving force, coming from the relaxation of line tension and, to a lesser extent, of twist (torsional) stress. The reason may be found in the persistence length of the free DNA, which is much longer (~50 nm) than the average radius of curvature in the nucleosome (~8 nm), and pushes the DNA to regain the straight average conformation on that length scale; the fact that spontaneous fluctuations were experimentally observed [89] both in presence of, and without binding proteins seems to support this view. In fact, our μs-long MD simulations were carried out with a soft restraining of the DNA linker (20 bp on each end), to simulate the effect of the background chromatin structure, and no fluctuations

larger than thermal vibrations were detected for the terminal phosphors; on the other hand, the DSB free end, once extended beyond a distance of about 2.5 nm away from the histone core, tended to regain a straight conformation and detach completely, confirming the importance of stress relaxation as a main driving force in DNA unwrapping. This might be the main force leading to spontaneous unwrapping of DSB cut ends, as well as of free nucleosome ends, opening the way to damage-signalling and repair proteins, and to remodelling factors, respectively. Notably, the important role observed for the mechanical stress suggests that such proteins should be implicated in complex mechanical actions on the nucleosome, resulting from the competition between internal stresses and chemical forces, very likely both sequence- and position-dependent.

8

CONCLUSIONS AND PERSPECTIVES

In this thesis I used single-molecule force spectroscopy performed with optical tweezers and molecular dynamics computer simulations, to study the structure, energetics, and dynamic evolution due to the occurrence of some specific defects in the DNA double helix. This is an important task, because a better understanding of the properties of lesions to the DNA structure at the molecular scale could provide a deeper knowledge of the microscopic effects of chemo- and radiation therapy, as well as improve our knowledge of many details of the repair mechanisms that cells activate to restore damaged genetic information.

The ability to detect point defects along the sequence of a DNA double strand is a problem of great relevance, both for molecular biology studies of DNA duplication and damage repair, and for the many technological applications exploiting the base-pair complementarity of nucleic acids. However, differences in base pairing between the native (Watson-Crick, A-T and G-C purine-pyrimidine pairs) and the defective sequence, for example including base mismatch defects (MM), are very subtle and may escape a direct experimental determination. The effect of a MM in a random position of a DNA sequence may reveal itself as a small local difference in free energy, which however shows up only when the two half-helices are split apart, e.g. during replication; or as a small variation in the elastic and mechanical response of the molecule to an applied force, e.g. by a repair enzyme.

In the experimental part of this work, I demonstrated the possibility to detect a single-base mismatch in short DNA hairpins, of 10 or 20 base-pairs in the stem, using the optical tweezers single-molecule force spectroscopy technique, in prof. Ritort's Small Systems Laboratory of the University of Barcelona. Equilibrium and out-of-equilibrium experiments were performed, with a different effectiveness depending on the type of hairpins. The equilibrium "hopping" experiments, realized by measuring the force fluctuations at fixed DNA extension, has proven to be more effective in the shorter hairpin sequences, because the fast kinetics rapidly reaches the equilibrium distribution; on the other hand, for the longer hairpin sequences the drift effects, due to the relaxation of the instrumental components at room temperature, often prevent attaining a stable equilibrium before the displacement of the optical trap can modify the equilibrium condition. In non-equilibrium "force-pulling" experiments, in which the optical trap position is moved at constant speed, the slower kinetic rate of the longer hairpins facilitates the detection of single jumps between folded and unfolded states. It is then easier to apply the theoretical considerations on the first-rupture force and extract information on the potential barriers and the free-energy of formation of

the mismatch defect. The shorter sequences during the pulling trajectories continuously jump between the unfolded and folded state, so that it is difficult to find a condition in which the initial and final state are perfectly defined.

In the optical tweezers experiments, I obtained important molecular-scale information on the free-energy of formation of the mismatch defects, the coexistence force, and the width of the energy barrier. The comparison of those properties among the different defects clearly indicates a variation in the mutated hairpins, compared with the native sequence. An underestimation of the measured free energy, compared to the theoretical values predicted by the nearest-neighbor (NN) model, was sometimes observed; this could be due to the extreme simplification of the potential barrier assumed for the unfolding transition using the Bell-Evans model, or to some inadequacy of the NN model for the particular geometry of the constrained hairpin. Such a circumstance leaves room to improve the model used to interpret the transition potential barrier, for example by using the Dudko-Hummer-Szabo model, based on Kramer's theory of transition states, and extract the ssDNA worm-like chain parameters, using a fit of the data with the Ξ-square error function as described in the thesis of A. Alemany [2]. A further improvement of the analysis reported in Chapter 3 could also be pursued, by extracting the free-energy profiles at any value of force, from the Boltzmann inversion already used in Section 3.6. While in that part of the analysis I showed the effect of the external force in deforming the energy landscape, the same data could be inverted to obtain the free-energy profile as a function of the hairpin opening at fixed force. Such data should be directly comparable to the predictions of the nearest-neighbor model.

Clearly, the amount of data obtained in this work is relatively limited, and a further extension of the experimental campaign should be envisaged, to improve the statistics, and to carry out similar measurements on different MM combinations.

A dedicated set of all-atom molecular dynamics simulations helped to understand microscopical phenomena that characterize the hairpin unfolding process, and suggested clues about the discrepancies observed in some experiments. In particular, it was observed that the first base-pair in the hairpin is constantly opened even under nearly zero force. This is due likely due to a pre-stress from the torsion of the DNA backbone. By accounting for this effect, a better agreement between experimental results and predictions of the NN model could be achieved. The timescale of MD simulation is anyway much faster than the experimental one; in this sense the MD results could be used as indication of the underlying physical processes, but could not replace the experiments. To reduce the gap between MD simulations and experiments one possible future development could be the simulation of the DNA hairpin with a coarse-grained MD model, such as the OxDNA [151], and use an implicit solvent description to accelerate the time-scale to conditions closer to experiments. With such a CG-MD model we could increase the timestep and very likely observe directly some hopping between the folded and unfolded state during

a simulated pulling trajectory. After such simulation, we could use the knowledge of each single energy contribution to the total energy barrier, to cure the apparent discrepancy observed in the application of the Crooks' fluctuation theorem to the experimental data.

In the second part of this thesis, I used MD simulations to study the properties of a different kind of defects in the DNA backbone, the single- and double-strand breaks. Cleary, such systems are complicate to study by force spectroscopy, since the type of damage could easily have a lifetime much shorter than the typical acquisition time of the optical tweezers. The first system I studied by MD simulation was the *linker DNA* connecting two nucleosomes in the 10-nm chromatin fiber. I simulated the evolution under external force and temperature, of radiation-induced strand breaks in a 31-bp DNA double helix. Simulated force spectroscopy revealed a rich dynamics, allowing not only to distinguish between different types of DNA strand breaks (single, or double with different structure), but also to learn a lot of important microscopic details. From the MD simulations, the absolute values of force necessary to break up a DSB-damaged DNA are quite large, of the order of 100 pN, at elongations of ~20%. Such values of longitudinal stress and strain are unlikely to be observed in the normal dynamics of chromatin, nor during chromosome mitosis. By comparison, thermal fluctuations seem unable to provide the energy necessary to overcome the barrier to rupture, unless the DSB is a very close one (i.e., the two breaks on opposite strands are separated by up to 2-3 base pairs at maximum). This allowed to rank the DSBs according to the width between the cuts, and attribute an average lifetime. I deduced that DSBs with spacing between the cut above 3 base-pairs should be stable over times much longer than the time-scale of repair protein action; therefore, only short-cut DSB seem to lead to complete fracture of the DNA, while wider-cut ones resist to both mechanical forces and temperature fluctuations. This finding may have an impact on the radiotherapy protocols and the treatment planning.

In a further development, I studied the evolution under external force and temperature of double-strand breaks (DSB) in nucleosomal DNA. I prepared a basic model of the nucleosome, made up of the canonical 8 histones plus a 187-bp DNA wrapped around. Using this configuration I inserted a DSB in four different positions on the DNA. The stability of the resulting structures was compared with the initial model. I observed that even the closely-cut DSBs remain relatively stable over long time scales, and display no sign of disassembly; interaction of the DSB ends with histone surfaces and tails is a main factor in damaged-DNA dynamics. DSB configurations close to histone tails in fact display a more active internal dynamics, with a participation also from histone fragment fluctuations.

Using the umbrella sampling methods I could show that the free-energy barriers for detachment of DNA from histones are relatively low, of the order of a few $k_B T$, implying that short sections of DNA could spontaneously unwrap over a time scale of >100 microseconds,

from DSB broken ends, or from the linker sections at the nucleosome ends, as observed in some experiments [89, 122, 160]. At the same time, histone tails represent a major steric obstacle for unwrapping of larger DNA sections, notwithstanding the driving force from stress relaxation. In fact, by performing a fully-consistent molecular stress calculations on the DNA wrapped in a nucleosome, it was revealed the existence of a strong internal driving force for straightening the circularly bent DNA. This force originates from the relaxation of line-tension and torsional stress injected in DNA by the wrapping around the histone core. I speculate that this might be the main force leading to spontaneous unwrapping of DSB cut ends, as well as of nucleosome ends, opening the way to damage-signalling and repair proteins, and to remodelling factors. Notably, such proteins should also be implicated in complex mechanical actions on the nucleosome, resulting from the competition between internal stresses and chemical forces, very likely both sequence- and position-dependent.

A detailed knowledge of the structural and mechanical response of DNA after radiation-induced damage is very relevant, since the repair machinery has a very high sensitivity to the strand break position and conformation, besides its dependence on the cell cycle phase. Given our simulation results, a key question is therefore: "what" the scouting proteins actually recognize at the damage site? For example, it has been postulated that, in the early stages of NHEJ, the Ku70/Ku80 heterodimer firstly binds to the open ends of cleaved DNA, on the basis of x-ray structures in which DNA fragments are co-crystallized with the protein monomers [127, 144]. However, this should be true only if we admit that DNA is firstly completely fractured into physically separate fragments. If we instead postulate, on the basis of the results of our MD simulations above, that DNA may not be fully broken, even after substantial radiation damage, could it be possible that such proteins have an even stronger affinity for some of the intermediate damaged states? Could such proteins actually identify *severely damaged*, rather than fully fractured DNA, and what the implications could then be, for the subsequent steps of the repair chain? Or, proteins as Ku70/80 recognize only *completely damaged* DNA and ignore other partially broken defects?

A necessary future development of these simulations will be the study of the interaction between DSB open ends, both in the linker and in the nucleosome, and typical signalization proteins, such as Ku70/80 for which PDB structures are already available.

An important implication of our findings is that DSBs actually undergoing spontaneous breaking by thermal fluctuations, should be only those with a short strand-break separation. DSBs in which the strand breaks are 3, 4 or more base-pairs apart can deform, indeed, and give rise to transient extremely distorted configurations; however, the DNA strands remain connected and maintain mechanical resistance. The MD simulations demonstrate that the π-stacking interaction can be strong enough to replace broken hydrogen bonds, in holding together the severely damaged DNA strands. Under external tension and torsion forces, instead, shorter DSBs split quite efficiently,

after the few hydrogen bonds holding the bases together are cleaved. On the other hand, the breaking of larger DSB proceeds by complex "stick/slip" sliding mechanisms, yet requiring substantially large forces and deformations. For the sake of completeness, it is worth noting that we did not consider at this stage of our study the possible clustering of DSBs, or their association with other types of defects, e.g. abasic sites, into more complex lesions. The eventual impact of such profound differences on, e.g., calibration of dose-response curves and tumor-control probability, to establish the efficiency of radiotherapy protocols, are open to further investigation.

BIBLIOGRAPHY

[1] N. C. Admal and E. Tadmor. "A unified interpretation of stress in molecular systems." In: *J. Elast.* 100.1-2 (2010), pp. 63–143.

[2] A. Alemany. *Dynamic force spectroscopy and folding kinetics in molecular systems*. 2014.

[3] A. Alemany and Ritort F. "Force-dependent folding and unfolding kinetics in DNA hairpins reveals transition-state displacements along a single pathway." In: *J Phys Chem Lett* 8 (2017), pp. 895–900.

[4] A. Alemany and F. Ritort. "Determination of the elastic properties of short ssDNA molecules by mechanically folding and unfolding DNA hairpins." In: *Biopolymers* 101 (2014), pp. 1193–1199.

[5] Mike P. Allen and Dominic J. Tildesley. *Computer simulation of liquids*. Oxford Science Publications, 1987.

[6] N. L. Allinger, Y. H. Yuh, and J. H. Lii. "Molecular mechanics. The MM3 force field for hydrocarbons." In: *J. Am. Chem. Soc.* 111 (1989), pp. 8551–8576.

[7] Andrea Amadei, Antonius B. M. Linnssen, and Hermann J. C. Berendsen. "Essential dynamics of proteins." In: *Proteins Str. Func. Gen.* 17 (1993), pp. 412–425.

[8] Brian R. Anderson, Julius Bogomolovas, Siegfried Labeit, and Henk Granzier. "Single-molecule force spectroscopy on titin implicates immunoglobulin domain stability as a cardiac disease mechanism." In: *J. Biol. Chem.* 288 (2013), pp. 5303–5315.

[9] A. Ashkin. "Acceleration and trapping of particles by radiation pressure." In: *Phys. Rev. Lett.* (1970), pp. 156–159.

[10] A. Ashkin. "Optical levitation by radiation pressure." In: *Appl. Phys. Lett.* (1971), pp. 283–285.

[11] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III." In: *J. Expt. Medicine* 79.2 (1944), pp. 137–158.

[12] Isabella Baccarelli, Ilko Bald, Franco A. Gianturco, Eugen Illenberger, and Janina Kopyra. "Electron-induced damage of {DNA} and its components: Experiments and theoretical models." In: *Phys. Rep.* 508.1-2 (2011), pp. 1–44.

[13] Joel S. Bader, Bruce J. Berne, and Eli Pollak. "Activated rate processes: The reactive flux method for onedimensional surface diffusion." In: *J. Chem. Phys.* 102 (1998), p. 4037.

[14] I. Bahar, T. R. Lezon, A. Bakan, and I. H. Shrivastava. "Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins." In: *Chem. Rev.* 110 (2010), pp. 1463–1497.

[15] Martina Banyay, Munna Sarkar, and Astrid Gräslund. "A library of IR bands of nucleic acids in solution." In: *Biophys. Chem.* 104.2 (2003), pp. 477–488.

[16] G. T. Barkema and N. Mousseau. "Event-based relaxation of continuous disordered systems." In: *Phys. Rev. Lett.* 77 (1996), p. 4358.

[17] George I. Bell. "Models for the specific adhesion of cells to cells." In: *Science* 200 (1978), pp. 618–627.

[18] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. "GROMACS: A message-passing parallel molecular dynamics implementation." In: *Comp. Phys. Comm* 91.1–3 (1995), pp. 43–56.

[19] R. B. Best, G. Hummer E. Paci, and O. K. Dudko. "Pulling direction as a reaction coordinate for the mechanical unfolding of single molecules." In: *The Journal of Physical Chemistry B* 112 (2008), pp. 5968–5976.

[20] C. Bouchiat, M. Wang, J.-f. Allemand, T. Strick, S. Block, and V. Croquette. "Estimating the persistence length of a worm-like chain molecule from force-extension measurements." In: *Biophys J* 76 (1999), pp. 409–413.

[21] B. D. Brower-Toland, C. L. Smith, R. C. Yeh, J. T. Lis, C. L. Peterson, and M. D. Wang. "Mechanical disruption of individual nucleosomes reveals a reversible multistage release of DNA." In: *Proc. Natl. Acad. Sci. USA* 99.4 (2002), pp. 1960–1965.

[22] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith. "Entropic elasticity of lambda-phage DNA." In: *Science* 265 (1994), pp. 1599–1601.

[23] Jean Cadet, Thierry Douki, and Jean-Luc Ravanat. "Oxidatively Generated Damage to the Guanine Moiety of DNA: Mechanistic Aspects and Formation in Cells." In: *Accounts Chem. Res.* 41.8 (2008), pp. 1075–1083.

[24] B. R. Cairns. "Chromatin remodeling: insights and intrigue from single molecule studies." In: *Nature Struct. Biol.* 14 (2007), pp. 989–996.

[25] J. Camunas-Soler, M. Ribezzi-Crivellari, and F. Ritort. "Elastic properties of nucleic acids by single-molecule force spectroscopy." In: *Annu. Rev. Biophys.* 45 (2016), pp. 65–84.

[26] W. J. Cannan and D. S. Pederson. "Mechanism and consequences of double-strand DNA break formation in chromatine." In: *J. Cell. Physiol.* (2016).

[27] Danilo Capecchi, Giuseppe Ruta, and Patrizia Trovalusci. "From classical to Voigt's molecular models in elasticity." In: *Arch. Hist. Exact Sci.* 64.5 (2010), pp. 525–559.

[28] Erwin Chargaff. "Some recent studies on the composition and structure of nucleic acids." In: *J. Cell Physiol. Suppl.* 38.1 (1951), pp. 41–59.

[29] F. Cleri. "Representation of mechanical loads in molecular dynamics simulations." In: *Phys. Rev. B* 65.1 (2001), p. 014107.

[30] S. Cocco, J. F. Marko, and R. Monasson. "Slow nucleic acid unzipping kinetics from sequence-defined barriers." In: *Eur PHys J E: Soft Matter Biol Phys* 10 (2003), pp. 153–161.

[31] A.R. Collins. "The comet assay for DNA damage and repair: principles, applications, and limitations." In: *Mol Biotechnol.* (2004), pp. 249–261.

[32] K. B. Connell, G. A. Horner, and S. Marqusee. "A single mutation at residue 25 populates the folding intermediate of E. coli RNase-H and reveals a highly dynamic partially folded ensemble." In: *J. Mol. Biol.* 391.2 (2009), pp. 461–470.

[33] J. Cormier, J. M. Rickman, and T. J. Delph. "Stress calculation in atomistic simulations of perfect and imperfect solids." In: *J. Appl. Phys.* 89 (2001), pp. 4198–4202.

[34] G. Crooks. "Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences." In: *Phys Rev E* 60 (1999), pp. 2721–2728.

[35] Gabor Csányi, Tristan Albaret, Mike C. Payne, and Alessandro De Vita. ""Learn on the Fly": a hybrid classical and quantum-mechanical molecular dynamics simulation." In: *Phys. Rev. Lett.* 93 (2004), p. 175503.

[36] Y. Cui and C. Bustamante. "Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure." In: *Proc. Natl. Acad. Sci. USA* 97 (2000), pp. 127–132.

[37] J. Dahm-Daphy, C. Sass, and W. Alberti. "Comparison of biological effects of DNA damage induced by ionizing radiation and hydrogen peroxide in CHO cells." In: *Int. J. Rad. Biol.* 76.1 (2000), pp. 67–75.

[38] I. Daidone and A. Amadei. "Essential dynamics: foundation and applications." In: *WIREs Comput Mol Sci* 2 (2012), pp. 762–770.

[39] Pablo D. Dans, Jürgen Walther, Hansel Gómez, and Modesto Orozco. "Multiscale simulation of DNA." In: *Curr. Opinion Struct. Biol.* 37 (2016), pp. 29–45.

[40] R. J. Davenport, G. J. L Wuite, R. Landick, and C. Bustamante. "ingle- molecule study of transcriptional pausing and arrest by E. coli RNA polymerase." In: *Science* 287 (2000), pp. 2497–2500.

[41] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 A resolution." In: *J. Mol. Biol.* 319 (2002), pp. 1097–1113.

[42] Grigory Dianov, Claus Bischoff, Jason Piotrowski, and Vilhelm A. Bohr. "Repair pathways for processing of 8-oxoguanine in DNA by mammalian cell extracts." In: *J. Biol. Chem.* 273 (1998), pp. 33811–33816.

[43] K. S. Dickson, C. M. Burns, and J. P. Richardson. "Determination of the free-energy change for repair of a DNA phosphodiester bond." In: *J. Biol. Chem.* 275.21 (2000), pp. 15828–15831.

[44] Christopher M. Dobson. "Experimental investigation of protein folding and misfolding." In: *Methods* 34.1 (2004), pp. 4–14.

[45] T. C. Doyle and J. L. Ericksen. "Nonlinear elasticity." In: *Advances in Applied Mechanics IV*. Ed. by H. L. Dryden and T. von Karman. Vol. 4. Elsevier, New York, 1956, pp. 53–115.

[46] Elise Dumont, Meilani Wibowo, Daniel Roca-Sanjuán, Marco Garavelli, Xavier Assfeld, and Antonio Monari. "Resolving the benzophenone DNA-photosensitization mechanism at QM/MM level." In: *J. Phys. Chem. Lett.* 6 (2015), pp. 576–580.

[47] M. Egli and Reinhard V. Gessner. "Stereoelectronic effects of deoxyribose O4' on DNA conformation." In: *Proc. Natl. Acad. Sci. USA* 92 (1995), pp. 180–184.

[48] Ramona Ettig, Nick Kepper, Rene Stehr, Gero Wedemann, and Karsten Rippe. "Dissecting DNA-histone interactions in the nucleosome by molecular dynamics simulations of DNA unwrapping." In: *Biophys. J.* 101 (2011), pp. 1999–2008.

[49] Evan Evans. "Probing the relation between force-lifetime-and chemistry in single molecular bonds." In: *Annu. Rev. Biophys. Biomol. Struct.* 30 (2001), pp. 105–128.

[50] Evan Evans and Ken Ritchie. "Dynamic strength of molecular adhesion bonds." In: *Biophys. J.* 72 (1997), pp. 1541–1555.

[51] N. Foloppe and A. D. MacKerell. "All-atom empirical force field for nucleic acids: 1) parameter optimization based on small molecule and condensed phase macromolecular target data." In: *J. Comp. Chem.* 21.2 (2000), pp. 86–104.

[52] N. R. Forde, D. Izhaky, G. R. Woodcock, G. J. Wuite, and C. Bustamante. "Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by Escherichia coli RNA polymerase." In: *Proc. Natl. Acad. Sci. USA* 99 (2002), pp. 11682–11687.

[53] N. Forns, S. de Lorenzo, M. Manosas, K. Hayashi, J. M. Huguet, and F. Ritort. "Improving signal/noise resolution in single-molecule experiments using molecular constructs with short handles." In: *Biophys J* 100 (2011), pp. 1765–1774.

[54] Daan Frenkel and Berend Smit. *Understanding molecular simulation*. Academic Press, 2002.

[55] R. Galindo-Murillo, D. R. Roe, and T. E. Cheatam. "Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC)." In: *Biochim. Biophys. Acta* 1850.5 (2015), pp. 1041–1058.

[56] Julian Garrec, Chandan Patel, Ursula Rothlisberger, and Elise Dumont. "Insights into intrastrand cross-link lesions of DNA from QM/MM Molecular Dynamics simulations." In: *J. Am. Chem. Soc.* 134.4 (2012), pp. 2111–2119.

[57] M. F. Goodman, S. Creighton, L. B. Bloom, and J. Petruska. "Biochemical basis of DNA replication fidelity." In: *J Crit Rev Biochem Mol Biol* 28 (1993), pp. 83–126.

[58] C. Gosse and V. Croquette. "Magnetic tweezers: micromanipulation and force measurement at the molecular level." In: *Biophys. J.* 82 (2002), pp. 3314–3329.

[59] G. Gouesbet and G. Grehan. "Generalized Lorenz–Mie theory for assemblies of spheres and aggregates." In: *J. Opt. A: Pure Appl Opt.* 1 (1999), pp. 706–712.

[60] Melville S. Green. "Markoff random processes and the statistical mechanics of timedependent phenomena." In: *J. Chem. Phys.* 22 (1954), p. 398.

[61] H. Grubmüller, B. Heymann, and P. Tavan. "Ligand binding and molecular mechanics calculation of the streptavidin-biotin rupture force." In: *Science* 271 (1996), pp. 997–999.

[62] Peter Hänggi, Peter Talkner, and Michael Borkovec. "Reaction-rate theory: fifty years after Kramers." In: *Rev. Mod. Phys.* 62.2 (1990), pp. 251–341.

[63] Steven Hayward and Bert L. de Groot. "Molecular modelling of proteins." In: vol. 443. Methods in molecular biology. Springer-Verlag, 2008. Chap. Normal Modes and Essential Dynamics, pp. 89–106.

[64] A. D. Hershey and M. Chase. "Independent functions of viral protein and nucleic acid in growth of bacteriophage." In: *J. Gen. Physiol.* 36.1 (1952), pp. 39–56.

[65] B. Hess, H. Bekker, H. J. C. Berendsen, and G. E. M. Fraaije. "LINCS: a linear constraint solver for molecular simulations." In: *J Comp Chem* 18 (1997), pp. 1463–1472.

[66] Jeff Hooyberghs, Paul Van Hummelen, and Enrico Carlon. "The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters"." In: *Nucl. Acids Res.* 37 (2009), e37.

[67] J. S. Hub, B. L. de Groot, and D. van der Spoel. "A free weighted histogram analysis implementation including robust error and autocorrelation estimates." In: *J. Chem. Theory Comput.* 6 (2010), pp. 3713–3720.

[68] G. Iliakis, H. Wang, A. R. Perrault, W. Boecker, B. Rosidi, F. Windhofer, W. Wu, J. Guan, G. Terzoudi, and G. Pantelias. "Mechanisms of DNA double strand break repair and chromosome aberration formation." In: *Cytogenet. Genome Res.* 104.1-4 (2004), pp. 14–20.

[69] J. H. Irving and J. G. Kirkwood. "The statistical mechanical theory of transport processes. IV. The equations of hydrodynamics." In: *J. Chem. Phys.* 18.6 (1950), pp. 817–829.

[70] Ken Ishikawa, Naofumi Handa, and Ichizo Kobayashi. "Cleavage of a model DNA replication fork by a Type I restriction endonuclease." In: *Nucl. Acids Res.* 37.11 (2009), pp. 3531–3544.

[71] B. Isralewitz, Mu Gao, and Klaus Schulten. "Steered molecular dynamics and mechanical functions of proteins." In: *Curr. Opinion Struct. Biol.* 11.2 (2001), pp. 224–230.

[72] B. Isralewitz, S. Izrailev, and K. Schulten. "Binding pathway of retinal to bacterio-opsin: a prediction by molecular dynamics simulations." In: *Biophys. J.* 73 (1997), pp. 2972–2979.

[73] Ivan Ivani et al. "Parmbsc1: a refined force field for DNA simulations." In: *Nature Meth.* 13 (2016), pp. 55–58.

[74] McCauley. M. J., L. Furman, C. A. Dietrich, I. Rouzina, M. E. Nunez, and M. C. Williams. "Quantifying the stability of oxidatively damaged DNA by single-molecule DNA stretching." In: *Nucl Acids Res* 46.8 (2018), pp. 4033–4043.

[75] Eva Jindrova, Stefanie Schmid-Nuoffer, Fabienne Hamburger, Pavel Janscak, and Thomas A. Bickle. "On the DNA cleavage mechanism of Type I restriction enzymes." In: *Nucl. Acids Res.* 33.6 (2005), pp. 1760–1766.

[76] H. Jónsson, G. Mills, and K. W. Jacobsen. "Nudged elastic band method for finding minimum energy paths of transitions." In: *Classical and Quantum Dynamics in Condensed Phase Simulations*. Ed. by B. J. Berne, G. Ciccotti, and D. F. Coker. World Scientific, 1998, p. 385.

[77] Ivan Junier, Alessandro Mossa, Maria Manosas, and Felix Ritort. "Recovery of free-energy branches in single molecule experiments." In: *Phys. Rev. Lett.* 102 (2009), p. 070602.

[78] Hiroo Kenzaki and Shoji Takada. "Partial unwrapping and histone tail dynamics in nucleosome revealed by coarse-grained molecular simulations." In: *PLOS Comp. Biol.* 11.8 (2014), e1004443.

[79] J. Kirkwood. "Statistical mechanics of fluid mixtures." In: *J. Chem. Phys.* 3 (1935), pp. 300–309.

[80] Mateusz Kogut, Cyprian Kleist, and Jacek Czub. "Molecular dynamics simulations reveal the balance of forces governing the formation of a guanine tetrad—a common structural unit of G-quadruplex DNA." In: *Nucl. Acids Res.* 44.7 (Apr. 2016), pp. 3020–3030.

[81] R. D. Kolodner. "Mismatch repair: mechanisms and relationship to cancer susceptibility." In: *Trends Biochem Sci* 20 (1995), pp. 397–401.

[82] A. H. Kramers. "Brownian motion in a field of force and the diffusion model of chemical reactions." In: *Physica* 7.4 (1940), pp. 284–304.

[83] Ryogo Kubo. "The fluctuation-dissipation theorem." In: *Rep. Progr. Phys.* 25 (1966), p. 255.

[84] Shankar Kumar, Djamal Bouzida, Robert H. Swendsen, Peter A. Kollman, and John M. Rosenberg. "The weighted histogram analysis method for free-energy calculations on biomolecules." In: *J. Comp. Chem.* 13.8 (1992), pp. 1011–1021.

[85] C. Kunz, Y. Saito, and P. Schär. "DNA Repair in mammalian cells. Mismatched repair: variations on a theme." In: *Cell Mol Life Sci* 66 (2009), pp. 1021–1038.

[86] A. Laio and M. Parrinello. "Escaping free-energy minima." In: *Proc. Natl. Acad. Sci. USA* 99.20 (2002), pp. 12562–12566.

[87] Brandon J. Lamarche, Nicole I. Orazio, and Matthew D. Weitzman. "The MRN complex in Double-Strand Break Repair and Telomere Maintenance." In: *FEBS Lett.* 584.17 (2010), pp. 3682–3695.

[88] Markita P. Landry, Patrick M. McCall, Zhi Qi, and Yann R. Chemla. "Characterization of photoactivated singlet oxygen damage in single-molecule optical trap experiments." In: *Biophys. J.* 97 (2009), pp. 2128–2136.

[89] Gu Li, Marcia Levitus, Carlos Bustamante, and Jonathan Widom. "Rapid spontaneous accessibility of nucleosomal DNA." In: *Nature Struct. Biol.* 12.1 (2005), pp. 46–53.

[90] M. R. Lieber. "The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway." In: *Annu. Rev. Biochem.* 79 (2010), pp. 181–211.

[91] Jehn-Huei Lii and Norman L. Allinger. "Directional hydrogen bonding in the MM3 force field." In: *J. Comput. Chem.* 19 (1998), pp. 1001–1016.

[92] Erik Lindahl, Berk Hess, and David van der Spoel. "GROMACS 3.0: a package for molecular simulation and trajectory analysis." In: *Molecular Modeling Annual* 7.8 (2001), pp. 306–317.

[93] M. Löbrich, B. Rydberg, and P. Cooper. "Repair of x-ray-induced DNA double-strand breaks in specific Not I restriction fragments in human fibroblasts: joining of correct and incorrect ends." In: *Proc. Natl. Acad. Sci. USA* 92 (1995), pp. 12050–12054.

[94] Raimo Lohikoski, Jussi Timonen, and Aatto Laaksonen. "Molecular dynamics simulation of single DNA stretching reveals a novel structure." In: *Chem. Phys. Lett.* 407.1-3 (2005), pp. 23–29.

[95] D. Lohr, R. Bash, H. Wang, J. Yodh, and S. M. Lindsay. "Using atomic force microscopy to study chromatin structure and nucleosome remodeling." In: *Methods* 41 (2007), pp. 333–341.

[96] J. F. Lutsko. "Stress and elastic constants in anisotropic solids: molecular dynamics techniques." In: *J. Appl. Phys.* 64 (1988), pp. 1152–1154.

[97] A. D. MacKerell and N. Banavali. "All-atom empirical force field for nucleic acids: 2) application to molecular dynamics simulations of DNA and RNA in solution." In: *J. Comp. Chem.* 21.2 (2000), pp. 105–120.

[98] A. D. MacKerell, M. Feig, and C. L. Brooks. "Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations." In: *J. Comp. Chem.* 25 (2004), pp. 1400–1415.

[99] A. D. MacKerell et al. "All-atom empirical potential for molecular modeling and dynamics studies of proteins." In: *J. Phys. Chem. B* 102 (1998), pp. 3586–3616.

[100] Andrew H. Mack, Daniel J. Schlingman, Robielyn P. Ilagan, Lynne Regan, and Simon G. J.Mochrie. "Kinetics and thermodynamics of phenotype: unwinding and rewinding the nucleosome." In: *J. Mol. Biol.* 423.5 (2012), pp. 687–701.

[101] F. Manca, S. Giordano, P. L. Palla, F. Cleri, and L. Colombo. "Two-state theory of single-molecule stretching experiments." In: *Phys. Rev. E* 87.3 (2013), p. 032705.

[102] Fabio Manca, Stefano Giordano, Pier Luca Palla, and Fabrizio Cleri. "Scaling shift in multicracked fiber bundles." In: *Phys. Rev. Lett.* 113.25 (2014), p. 255501.

[103] Fabio Manca, Stefano Giordano, Pier Luca Palla, and Fabrizio Cleri. "Stochastic mechanical degradation of multi-cracked fiber bundles with elastic and viscous interactions." In: *Eur. Phys. J. E* 38.5 (2015), pp. 1–21.

[104] M. Manosas, D. Collin, and F. Ritort. "Force-dependent fragility in RNA hairpins." In: *Phys Rev Lett* 96 (2006), p. 218301.

[105] Vasilissa Manova and Damian Gruzska. "DNA damage and repair in plants - from models to crops." In: *Front. Plant Sci.* 6 (2015), p. 885.

[106] J. F. Marko and E. D. Siggia. "Statistical mechanics of supercoiled DNA." In: *Phys. Rev. E* 52 (1995), pp. 2912–2938.

[107] Robert K. McGinty and Song Tan. "Nucleosome Structure and Function." In: *Chem. Rev.* 115.6 (2015), pp. 2255–2273.

[108] A. Meyerhans and J.-P. Vartanian. "The fidelity of cellular and viral polymerases and its manipulation." In: *Origin and evolution of viruses*. Ed. by E. Domingo, R. G. Webster, and J. F. Holland. New York: Academic Press, 1999. Chap. 5.

[109] L. Mistura. "The definition of the pressure tensor in the statistical mechanics of nonuniform classical fluids." In: *Int. J. Thermophys.* 8.3 (1987), pp. 397–403.

[110] P. Modrich. "DNA mismatch correction." In: *Annu Rev Biochem* 56 (1987), pp. 435–466.

[111] P. Modrich. "Mechanisms in eukaryotic mismatch repair." In: *J Biol Chem* 281 (2006), pp. 30305–30309.

[112] G. Morrison, C. Hyeon, M. Hinczewski, and D. Thirumalai. "Compaction and tensile forces determine the accuracy of folding landscape parameters from single molecule pulling experiments." In: *Physical Review Letters* 106 (2011).

[113] S. Nandhakumar, S. Parasuraman, M. M. Shanmugam, K. Ramachandra Rao, Parkash Chand, and B. Vishnu Bhat. "Evaluation of DNA damage using single-cell gel electrophoresis (Comet Assay)." In: *J Pharmacol Pharmacother.* (2011), pp. 107–111.

[114] Karl C. Neuman and Steven M. Block. "Optical trapping." In: *Rev. Sci. Instrum.* 75.9 (2004), pp. 2787–2809.

[115] B. Ogorek and P. E. Bryant. "Repair of DNA single-strand breaks in X-irradiated yeast. II. Kinetics of repair as measured by the DNA-unwinding method." In: *Mutat. Res.* 146.1 (1985), pp. 63–70.

[116] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco. "Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers AMBER force field for nucleic acids: improving the description of alpha/gamma conformers." In: *Biophys. J.* 92.11 (2007), pp. 3817–3829.

[117] Gregoire Perret et al. "Real-time mechanical characterization of DNA degradation under therapeutic x-ray and its theoretical modelling." In: *Microsyst. Nanoeng.* (2016).

[118] John Petruska and Myron F. Goodman. "Enthalpy-Entropy Compensation in DNA Melting Thermodynamics." In: *J. Biol. Chem.* 270 (1995), pp. 746–750.

[119] N. Peyret, P. A. Seneviratne, H.T. Allawi, and J. SantaLucia. "Nearest-neighbor thermodynamics and NMR of DNA sequences with internal AA, CC, GG, and TT mismatches." In: *Biochemistry* 38 (1999), pp. 3468–3477.

[120] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kale, and Klaus Schulten. "Scalable molecular dynamics with NAMD." In: *J. Comp. Chem.* 26 (2005), pp. 1781–1802.

[121] L. H. Pope, M. L. Bennink, K. A. van Leijenhorst-Groener, D. Nikova, J. Greve, and J.F. Marko. "Single chromatin fiber stretching reveals physically distinct populations of disassembly events." In: *Biophys. J.* 88 (2005), pp. 3572–3583.

[122] R. U. Protacio, K. J. Polach, and J. Widom. "Coupled-enzymatic assays for the rate and mechanism of DNA site exposure in a nucleosome." In: *J. Mol. Biol.* 274 (1997), pp. 708–721.

[123] Dennis C. Rapaport. *The Art of Molecular Dynamics Simulation*. Cambridge University Press, 2004.

[124] Swarnalatha Y. Reddy, Fabrice Leclerc, and Martin Karplus. "DNA Polymorphism: A Comparison of Force Fields for Nucleic Acids." In: *Biophys. J.* 84.3 (2003), pp. 1421–1449.

[125] E. Reynaud. "Protein Misfolding and Degenerative Diseases." In: *Nature Education* (2010).

[126] F. Ritort. "Single-molecule experiments in biological physics: methods and applications." In: *J Phys Condens Matter* 18 (2006), R531–R583.

[127] A. Rivera-Calzada, L. Spagnolo, L. H. Pearl, and O. Llorca. "Structural model of full-length human Ku70-Ku80 heterodimer and its recognition of DNA and DNA-PKcs." In: *EMBO Rep.* 8.1 (2007), pp. 56–62.

[128] G. Rossetti, P. D. Dans, I. Gomez-Pinto, I. Ivani, C. Gonzalez, and M. Orozco. "The structural impact of DNA mismatches." In: *Nucl Acids Res* 43 (2015), pp. 4309–4321.

[129] Léon Sanche. "Beyond radical thinking." In: *Nature* 461 (2009), pp. 358–359.

[130] John SantaLucia. "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics." In: *Proc. Natl. Acad. Sci. USA* 95.4 (1998), pp. 1460–1465.

[131] John SantaLucia and Donald Hicks. "The thermodynamics of DNA structural motifs." In: *Annu. Rev. Biophys. Biomol. Struct.* 33 (2004), pp. 415–440.

[132] Mogurampelly Santosh and Prabal K Maiti. "Force induced DNA melting." In: *J. Phys.: Cond. Matt.* 21.3 (2009), p. 034113.

[133] Masahiko S. Satoh and Tomas Lindahl. "Role of poly(ADP-ribose) formation in DNA repair." In: *Nature* 356 (1992), pp. 356–358.

[134] A. Savelyev and A. D. MacKerell. "All-atom polarizable force field for DNA based on the classical Drude oscillator model." In: *J. Comp. Chem.* 35 (2014), pp. 1219–1239.

[135] Jürgen Schlitter. "Estimation of absolute and relative entropies of macromolecules using the covariance matrix." In: *Chem. Phys. Lett.* 215.6 (1993), pp. 617–621.

[136] E. A. Shank, C. Cecconi, J. W. Dill, S. Marqusee, and C. Bustamante. "The folding cooperativity of a protein is controlled by its chain topology." In: *Nature* 465 (2010), pp. 637–640.

[137] A.K. Shaytan, G.A. Armeev, A. Goncearenco, V.B. Zhurkin, D. Landsman, and A. R. Panchenko. "Coupling between histone conformations and DNA geometry in nucleosomes on a microsecond timescale: atomistic insights into nucleosome functions." In: *J. Mol. Biol.* 428 (2016), pp. 221–237.

[138] Katrin R. Siefermann, Yaxing Liu, Evgeny Lugovoy, Oliver Link, Manfred Faubel, Udo Buck, Bernd Winter, and Bernd Abel. "Binding energies, lifetimes and implications of bulk and interface solvated electrons in water." In: *Nature Chem.* 2.4 (2010), pp. 274–279.

[139] Raghvendra Pratap Singh, Ralf Blossey, and Fabrizio Cleri. "Structure and mechanical characterization of DNA i-Motif nanowires by molecular dynamics simulation." In: *Biophys. J.* 105.12 (2013), pp. 2820–2831.

[140] L. Skjaerven, S. M. Hollup, and Nathalie Reuter. "Normal mode analysis for proteins." In: *Journal of Molecular Structure: THEOCHEM* 898 (2009), pp. 42–48.

[141] S B Smith, Y Cui, and C Bustamante. "Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules." In: *Science* 271 (1996), pp. 795–799.

[142] Clemens von Sonntag. *Free-radical-induced DNA Damage and its Repair. A Chemical Perspective*. Berlin Heidelberg: Springer-Verlag, 2006.

[143] M. R. Sorensen and A. F. Voter. "Temperature-accelerated dynamics for simulation of infrequent events." In: *J. Chem. Phys.* 112.21 (2000), pp. 9599–9606.

[144] A. Spagnolo, A. Rivera-Calzada, L. H. Pearl, and O. Llorca. "Three-Dimensional Structure of the Human DNA-PKcs/Ku70/Ku80 Complex Assembled on DNA and Its Implications for DNA DSB Repair." In: *Mol. Cell* 22.4 (2006), pp. 511–519.

[145] Katrin Spiegel and Alessandra Magistrato. "Modeling anticancer drug–DNA interactions via mixed QM/MM molecular dynamics simulations." In: *Org. Biomol. Chem.* 4 (2006), pp. 2507–2517.

[146] Justin Spiriti, Hiqmet Kamberaj, Adam M. R. de Graff, M. F. Thorpe, and Arjan van der Vaart. "DNA Bending through Large Angles Is Aided by Ionic Screening." In: *J. Chem. Theory Comput.* 8.6 (2012), pp. 2145–2156.

[147] Michiel Sprik and Giovanni Ciccoti. "Free energy from constrained molecular dynamics." In: *J. Chem. Phys.* 109 (1998), p. 7737.

[148] T R Strick, J F Allemand, D Bensimon, A Bensimon, and V Croquette. "The elasticity of a single supercoiled DNA molecule." In: *Science* 271 (1996), pp. 1835–1837.

[149] T. Strick, J. F. Allemand, V. Croquette, and D. Bensimon. "Twisting and stretching single DNA molecules." In: *Progr. Biophys. Mol. Biol.* 74.1-2 (2000), pp. 115–140.

[150] Arun K. Subramaniyan and C. T. Sun. "Continuum interpretation of virial stress in molecular simulations." In: *Int. J. Solids Struct.* 45 (2008), pp. 4340–4346.

[151] P. Sulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis. "Sequence-dependent thermodynamics of a coarse-grained DNA model." In: *J. Chem. Phys.* 135.13 (2012), p. 135101.

[152] Y. Suzuki and O. K. Dudko. "Single-molecule rupture dynamics on multidimensional landscapes." In: *Phys Rev Lett* 104 (2010), p. 048101.

[153] Halina Szatyłowicz and Nina Sadlej-Sosnowska. "Characterizing the Strength of Individual Hydrogen Bonds in DNA Base Pairs." In: *J. Chem. Inf. Mod.* 50.12 (Dec. 2010), pp. 2151–2161.

[154] E. Tadmor and R. E. Miller. *Modeling materials: continuum, atomistic and multiscale techniques*. Cambridge University Press, Cambridge, 2011.

[155] R. Tice and R. Setlow. "Handbook of the biology of aging." In: *Handook of the biology of aging*. Ed. by C. E. Finch and E. L. Schneider. New York: Van Nostrand Reinhold, 1985.

[156] A. Tikhomirova, I. V. Beletskaya, and T. V. Chalikian. "Stability of DNA duplexes containing GG, CC, AA, and TT mismatches." In: *Biochemistry* 45 (2006), pp. 10563–10571.

[157] Alejandro Torres-Sanchez, Juan M. Vanegas, and Marino Arroyo. "Examining the mechanical equilibrium of microscopic stresses in molecular simulations." In: *Phys. Rev. Lett.* 114 (2015), p. 258102.

[158] Alejandro Torres-Sanchez, Juan M. Vanegas, and Marino Arroyo. "Geometric derivation of the microscopic stress: A covariant central force decomposition." In: *J. Mech. Phys. Solids* 93 (2016), pp. 224–239.

[159] G. M. Torrie and J. P. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling." In: *J. Comp. Phys.* 23.2 (1977), pp. 187–199.

[160] Katalin Tóth, Vera Böhm, Carolin Sellmann, Maria Danner, Janina Hanne, Marina Berg, Ina Barz, Alexander Gansen, and Jörg Langowski. "Histone- and DNA-sequence-dependent stability of nucleosomes studied by single-pair FRET." In: *Cytometry* A83 (2013), pp. 839–846.

[161] M. S. Vijayabaskar. "Introduction to hidden Markov models and its applications in biology." In: *Meth Mol Biol* 1552 (2017), pp. 1–12.

[162] Stephane Vispé and Masahiko S. Satoh. "DNA repair patch-mediated double strand DNA break formation in human cells." In: *J. Biol. Chem.* 275.35 (2000), pp. 27386–27392.

[163] Karine Voltz, Joanna Trylska, Nicolas Calimet, Jeremy C. Smith, and Jörg Langowski. "Unwrapping of nucleosomal DNA ends: a multiscale molecular dynamics study." In: *Biophys. J.* 102 (2012), pp. 849–858.

[164] A. F. Voter and M. R. Sorensen. "Accelerating atomistic simulation of defect dynamics." In: *MRS Proceedings*. Vol. 538. 1999, pp. 427–439.

[165] Arthur Voter. "A method for accelerating the molecular dynamics simulation of infrequent events." In: *J. Chem. Phys.* 106.1 (1997), pp. 4665–4667.

[166] C.-C. Wang, K. Sivashanmugan, C.-K. Chen, J.-R. Hong, W.-I Sung, J.-D. Liao, and Y.-S. Yang. "Specific unbinding forces between mutated human P-selectin glycoprotein Ligand-1 and viral Protein-1 measured using force spectroscopy." In: *J. Phys. Chem. Lett.* 8.21 (2017), pp. 5290–5295.

[167] J. F. Ward. "DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability." In: *Prog. Nucl. Acids Res.* 35 (1988), pp. 95–125.

[168] J. Wildenberg and M. Meselson. "Mismatch repair in heteroduplex DNA." In: *Proc Natl Acad Sci USA* 72 (1975), pp. 2202–2206.

[169] Claire Wyman and Roland Kanaar. "DNA Double-Strand Break Repair: All's Well that Ends Well." In: *Annu. Rev. Genetics* 40 (2006), pp. 363–383.

[170] Jie Yan, Thomas J. Maresca, Dunja Skoko, Christian D. Adams, Botao Xiao, Morten O. Christensen, Rebecca Heald, and John F. Marko. "Micromanipulation Studies of Chromatin Fibers in Xenopus Egg Extracts Reveal ATP-dependent Chromatin Assembly Dynamics." In: *Mol. Biol. Cell.* 18.2 (2007), pp. 464–474.

[171] Jan Řezáč, Pavel Hobza, and Sarah A. Harris. "Stretched DNA investigated using molecular dynamics and quantum mechanical calculations." In: *Biophys. J.* 98.1 (2010), pp. 101–110.

APPENDICES

A

SOME DETAILS ON THE MINI-TWEEZER APPARATUS

A.1 OPTICAL PATH

The lasers used to generate the two beams have a wavelength of 845 nm, a power of 200 mW, and produce the fundamental transverse electromagnetic mode, linearly polarized, with a Gaussian profile. The laser frequency has been chosen in order to reduce the water absorption and possible electronic level interactions in the biomolecules. The intensity of the laser beams and the temperature of the laser diodes are controlled by an independent power supply; normally they are used at a lower power than the nominal one. The two lasers follow symmetric paths that cross inside the microfluidic chamber, and share a common optical axis. In this way, with a correct calibration of the laser power, the two parallel contributions to the force applied on the trapped particle can cancel each other out. The optical scheme of the mini-tweezer is shown in Figure 85.

The light intensity generated by each laser is funnelled with a single-mode optical fiber, and filtered to produce a sharper wavelength profile. The other fiber extremity not connected with the laser source is attached to a *wiggler*. The wiggler uses two perpendicular piezoelectric crystals to modify the direction of the laser beam. The modification is actuated and controlled by the electronic controller connected to the computer. This setup allows to change the laser path and displace the position of the trap inside the chamber, via the host software installed on the same computer; moreover, specific protocols for the software (such as the constant-pulling velocity, or the constant -force) can be implemented to facilitate or automate the experiments.

Part of the the light emerging from wiggler (about 8%) is split by using a pellicle beam-splitter, to form a *light-lever*. The position of the light-lever beam, after having been refocused by an aspherical lens, is measured by a position-sensitive detector (PSD). In the PSD, the displacement of a light spot on the sensitive surface of the detector is transformed into an analog current output, read by the electronic controller, so the planar position of the trap (x, y) can be measured. The two coordinates individuate the position in the plane orthogonal to the light propagation direction and, as said, could be varied during the experiments, while the longitudinal position (parallel to the optical axis) is fixed by the focal point of the microscope objectives that form the trap. With this information the spatial position of the trap inside the chamber can be fixed and kept under control.

The remaining light intensity, before entering the objective that shapes the trap, is collimated, aligned with the optical axis, and transformed to a circularly-polarized light beam. These passages are done by using different devices, respectively: a planar lens; a *polarizing*

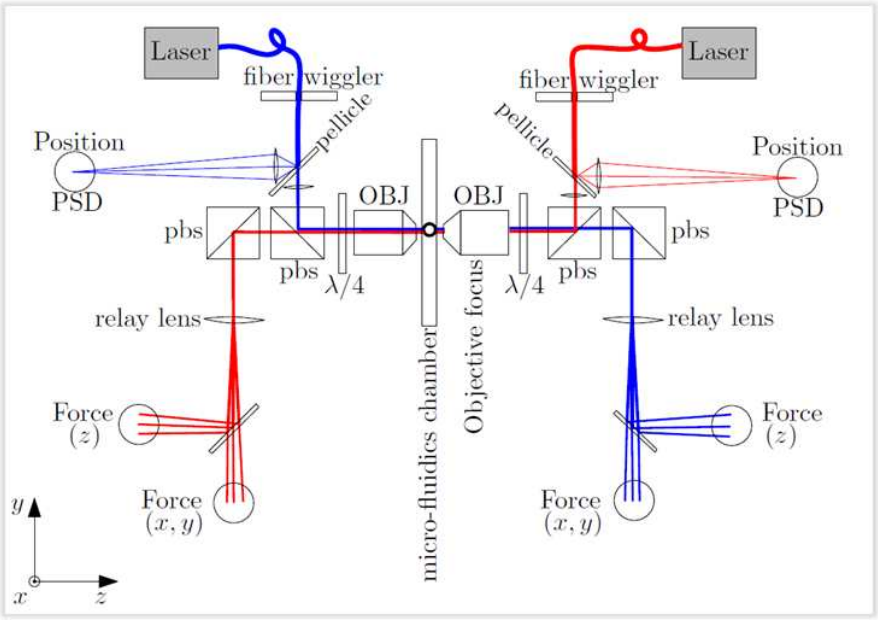


Figure 85: Scheme of the dual-laser mini-tweezer. In evidence (red and blue) the optical path of the two lasers. The light intensity generated by the laser sources are funnelled by the optical fiber to the wiggler, where a piezoelectric motor allows to modify the beam direction and, as a direct consequence, displace the position of the optical trap inside the chamber. Then, the lasers pass through the pellicle film where each one gets split in two: one beam is focused on a PSD detector and will be used to determine the position of the trap; the other beam continues its path and after some passages arrives at the objective, where a lens focus the beam into the microfluidic chamber to form the trap. After the interaction, the outgoing beam is collected by the opposite objective and directed on two sensors that will measure the components of the force applied on the trapped particle.

beam-splitter (PBS) that also selects the horizontal polarization; and a quarter-wave plate. The circular polarization ensures that the force exerted on the trapped particle will not depend on the polarization of the incident light.

The objective used to focus the beam is a standard microscope objective of the *water-immersion* type, with a numerical aperture of 1.2; such wide aperture is necessary in order to obtain a sharp and concentrated laser profile in the $\hat{z}$ direction (Figure 86).

From the objective lens, each beam enters the microfluidic chamber and can interact with the particle (usually a silica bead of radius a few μm). Each outgoing laser beam is collected by the opposite objective and sent to another quarter-wave plate, where it is converted to vertically-polarized light. Since the polarization is orthogonal to that of the counter-propagating laser beam, it is possible to separate the two paths with the same PBS used for the incoming symmetric beam. The outgoing beam continues its straight propagation, arriving on a second PSD that will extract it from the optical axis and, after a passage on a relay lens, is split in two. Each half-beam is sent to a different sensor: a PSD that measures the displacement from

the rest position, to get information on the orthogonal components (F_x, F_y) of the force applied on the trapped object; and a "bullseye" filter followed by a photo-diode, designed to measure the parallel force component F_z, along the beam propagation axis.

With this setup positions and forces are collected but, as an ulterior check for a direct control of what is happening inside the microfluidic chamber, an imaging system is introduced. A blue *LED* (wavelength 470 nm) is expanded about the optical axis with a lens, thus passing through the PBSs, the objectives and uniformly illuminates the focal plane of the laser beam; then it is captured by the opposite objective, and again passing through the PBSs it is separated by the lasers used for the force measurements, and projected on a *CCD camera* by a lens. The camera detects both near-infrared and visible light, to obtain a real-time image of the experiment. To not be blinded by the intense light emission of the lasers, a filter is applied before entering the CCD (in special conditions, like during the initial alignment of the focal point of the two beams, this filter can be removed to detect the spot position; clearly, this must be done at a reduced laser intensity).

A.2 CHAMBER PREPARATION

The microfluidic chamber is constituted by two rectangular glass slides (24 x 60 mm) separated by a parafilm layer in which three channels are printed. One of the glass coverslips carries six holes (three on each short side) matching the tube connections on the support. These are the connection inlet/outlet from which the solution can be flown inside the chamber, or collected to the trash. The holes are created using a laser printer after repetitive cycles of impression on the same glass (depending on the power it could be necessary around 5-10 shots). A similar technique is used to cut the parafilm layer to print the shape of the fluidic chamber. In the standard design three linear channels not connected to each other are realized (Figure 87). The lateral channels are used to flow the beads used in the experiment; they are connected to the central one by two quartz microtube. A third quartz tube, the micropipette, is inserted after the filler tube end (the direction is determined by fluid flow), and it is designed with a thin tip to be used during the experiments to capture the streptavidine beads, by air suction.

The micropipette is fabricated with the *pipette puller*, a simple device that melts a quartz tube under stretching. The melted area when pulled elongates and reduces the section of the tube, giving a micron-thin tip. This end is to be put inside the chamber, while the other (wider) end is connected to a syringe (usually 1.0 ml) by a plastic micro-tube, to create the negative pressure needed to capture the beads.

The chambers are produced directly on site before each experiment. The procedure to build a chamber requires: (i) to take a glass slide without holes, (ii) stick a parafilm layer on it, then (iii) put in the correct position the two filler tubes and the micro-pipette, (iv) sandwich

them between another parafilm layer, and finally (v) add the coverslip glass with the holes. To fix the two layers they must be warmed with a heater ($\sim$ 120 °C), taking care to not melt also the channel structure, nor to obstruct the holes.

Once the chamber is ready it is mounted on the support, aligning the holes with the supply tubes, and verifying that the channels, the filler tube and the micropipette work correctly. Then, the support is inserted between the two microscope objectives of the mini-tweezer, where it is held by a motorized-xyz stage with fine positioning control ($< 0.5\mu$m). However, the motorized support is used only during the preparatory stage of the experiment to capture the beads. Instead, during the single-molecule experiments the microfluidic chamber is kept fixed and the trap position is moved inside the chamber.

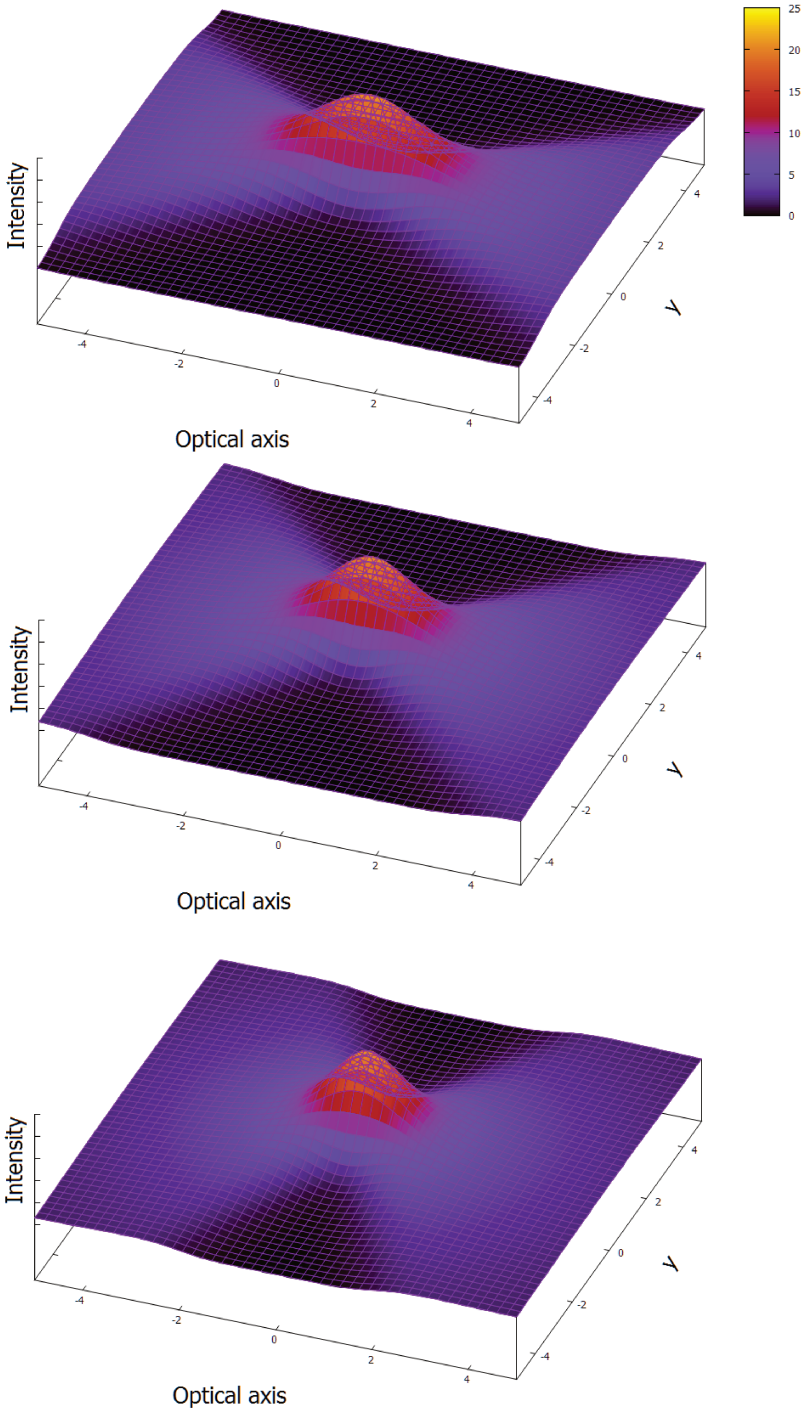


Figure 86: Laser intensity profile at different numerical aperture. From left to right three different schematic representatio of the Intensity profile at increasing numerical aperture. Higher numerical aperture mean a sharper intensity profile along the optical axis and as consequence a stronger attractive force to the maximum.

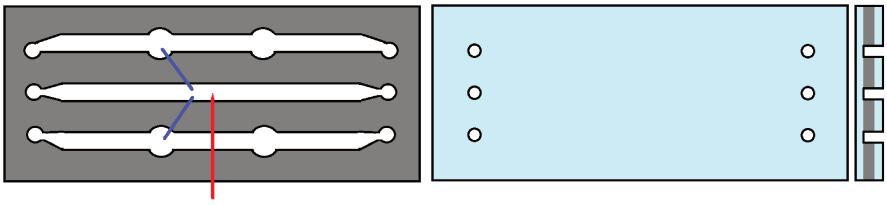


Figure 87: Microfluidic chamber schema. On the left the parafilm design with three linear channels: the two lateral are connected via quartz microtubes (*filler tube* blue) to the central one, where the experiments actually take place. Close to the end of the filler tube is inserted the micropipette (red), which is used to capture the streptavidine-coated bead. In the center, a schematic of the glass coverslip with the six holes. At the extreme right, a cross-section of the chamber shows the three layers, with the parafilm sandwiched between the two glass slides and the holes of the coverslip are aligned with the channel termination.

B

HIDDEN MARKOV MODEL

The Hidden Markov Model (HMM) is a probabilistic technique used in the study of a time series of data. Given a discrete time-series of observations $\hat{O} = (o_1, o_2, \ldots, o_T)$, the HMM assigns a likelihood to a parametric (λ) probability model, and iteratively adjusts the parameter to increase the likelihood of the model with respect to the series of data. The output of the method is a finite-state Markov chain $\hat{S} = (s_1, s_2, \ldots, s_T)$, and a finite set of output probabilities distribution $\{f_i^\lambda(o)\}$, $\forall o \in \hat{O}$, the probability that an observation o corresponds to the state $n = 1 \ldots N$ being $f_n^\lambda(o) \in [0, 1]$.

A Markov process has no memory so the total probability to observe a state n at time t_i is:

$$P_n(t_i) = \sum_m T_{nm} P_m(t_{i-1}) \tag{B.1}$$

where T_{nm} is the transition matrix, which is assumed to not vary with time. The Markov chain is a *path* representing the succession of states that have the maximum probability of describing the current observation. The adjective "hidden" in the name came from the veil imposed by the model on the series of observations: actually, from a time-serie spanning a continuous space of values, the HMM extracts the *path*, a succession of indices indicating the state in which the system is most probably to be found (Figures 88,89 top).

We used the Baum-Welch algorithm [161] to implement the maximum likelihood re-estimator for a single time-series observation. To this end, we tentatively assumed N states (fixed by the user) with Gaussian distribution, identified by three independent parameters:

- average value in the state n: μ_n
- variance about the state: σ_n
- initial probability for the state: P_n^{init}

As a consequence, the HMM assigns to each observation a probability of corresponding to a "hidden" state, given by:

$$f_n(o) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{o-\mu_n}{\sigma_n}\right)^2\right] \tag{B.2}$$

The software has been written in JAVA to allow compatibility on differents operating systems. The program accept as input :

-N **NumberOfStates** Number of states use in the HiddenMarkov Model.

-P **StartingProbabilities** List of N double that describe the probability of each state in the starting step (not mandatory if not set the algorithm will assign equal probability to all state).

-s **GaussianStatesParameters** List of 2xN double that describe the Gaussian probability (mu_1,σ_1,μ_2,$\sigma_2 \ldots \mu_n$,σ_n).

-t **TransitionMatrix** List of NxN double that will be arrange in a NxN matrix, T_{ij}=List[i/N][j%N].

-f **Input file** String that contains the full path name of the input file.

-c **ColumnIndex** Column of the file containing the time-serie.

-o **Output file** The path of the file where the data and the best trajectory are written.

-xml **XML OutputFile** Path of the XML file containing the information about all the iterations.

Other, not mandatory parameters needed to modify the criteria of convergence in the iterative algorithm are:

-iter **MaxIteration** Maximum number of iteration in the Bauman-Welch's optimization algorithm.

-conv **OptimizationConvergence** Convergence parameter, if it's set negative the criteria won't be used and only *MaxIteration*.

-err **NormalizationError** Max accepted error in the normalization of forward and backward coefficients.

Furthermore, it is possible to automatically produce a plot of the time-series to be compared to the Markov-model path, and of the model-fitted parameters to be compared to the data histogram with the help of some additional parameters (-gpl, -hist,...).

Analyzing the results of force spectroscopy, we look at a time-series of states changing between two states ("folded" and "unfolded") and all the possible intermediate states between these two. When the force jump between the two extreme states is larger than the thermal noise, collecting information on the instantaneous force is equivalent to collecting information on the occupation of each state (e.g., folded vs. unfolded state of the hairpin). Close to the coexistence value of the parameter λ_c, at which the two states have equal occupation probability, the kinetic rate of unfolding (k^+) and refolding (k^-) fall in a timescale that allows to observe several hopping events (Fig.25). Therefore, for each time-series taken at fixed λ, the HMM can reconstruct:

(i) the probability distribution of finding the hairpin in the folded/unfolded state (see histogram in Fig.25);

(ii) the average force in the folded/unfolded state;

(iii) the sequence of transitions along the time evolution (the green line in the time series Fig. 25);

(iv) the transition probability between the two states (directly related to the reaction kinetic coefficients k^+, k^-).

Each variation of λ produces a consequent variation of the probability to observe the system in one of the two states, such as the folded (w_N) or unfolded (w_U) state. Upon repeating the measurements for a collection of time-series taken at different λ's, one can measure the relative variation of the occupation probability of the two states. Typical time-series and probability histogram data for 10bp hairpins, with native sequence or including a GA and GT mismatch, are presented in the multi-panel Figures 26, 27 and 28.

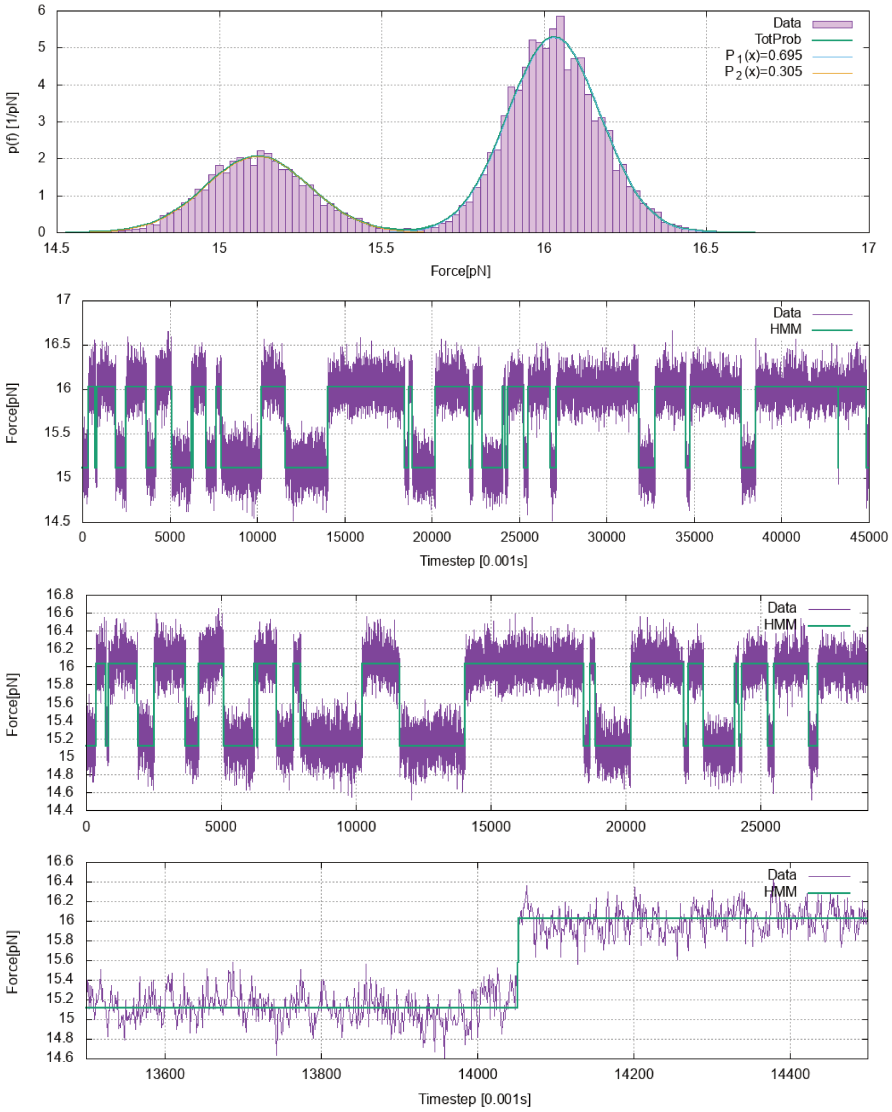


Figure 88: Example of a time-series of force measurements (center) from a hopping experiment of the 20bp native hairpin; the time-series represents the recorded force as a function of the instrument time step (dt = 0.001s). On the bottom plot a zoom of the same time-series for small interval of time around a jump. The plot above is the histogram of the collected values of instantaneous force, together with the best-fit from the Hidden-Markov model (green curve); the two Gaussian peaks correspond to the force distributions in the folded and unfolded state, in the legend are reported the relative probabilities.

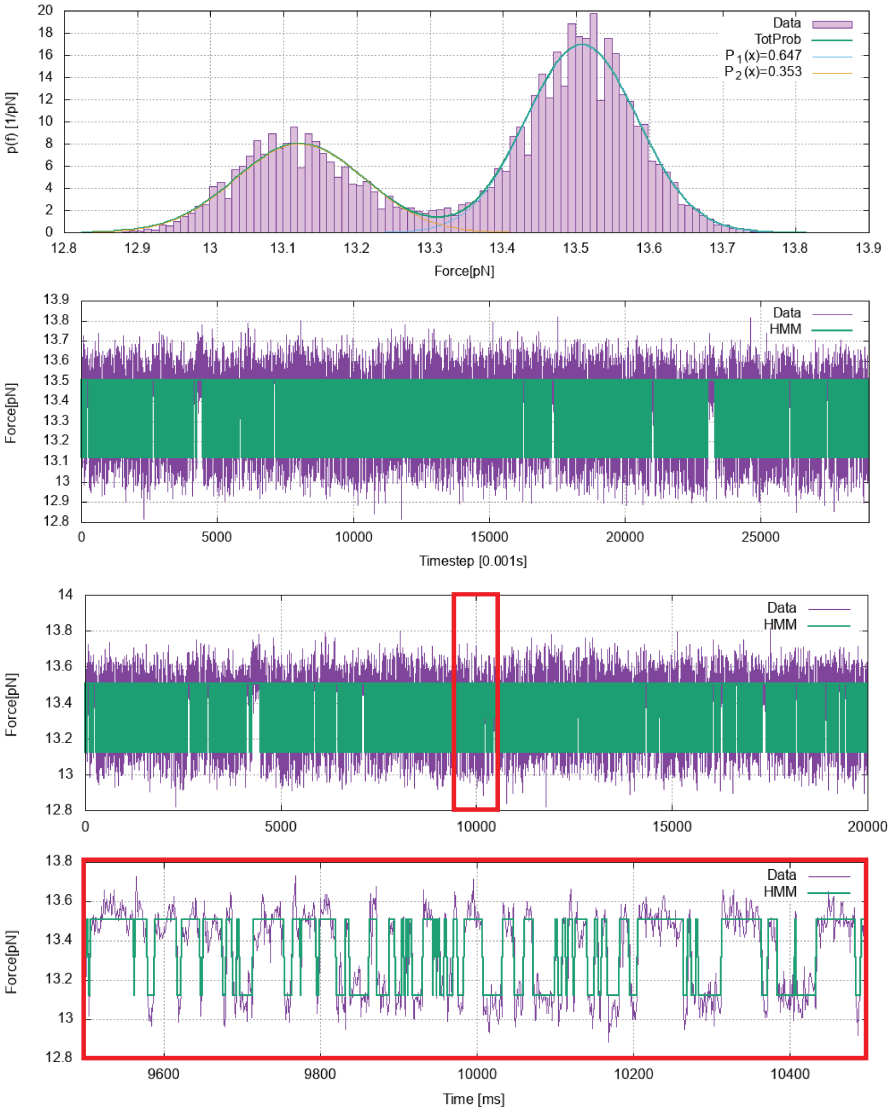


Figure 89: Time-series of force measurements from a hopping experiments, analogue to Fig.88 of the 10bp native hairpin.

C

ALIGNING TRAJECTORIES IN THE OPTICAL TRAP

The non-equilibrium pulling experiments are performed by moving the position of the optical trap inside the microfluidic chamber. When a molecule is bound between the fixed bead and the bead in the trap, the displacement of the latter results in the stretching of the molecule. Repeating cycles of extension and contraction of the molecule, one can measure important out-of-equilibrium properties such as the first-rupture force, the average work, and so on, from the relation between the applied force and the extension measured.

While such measurements are collected, the variation of distance is assumed to be controlled by the displacement measured on the light-lever; however, some mechanical relaxation of the chamber or minor movements of the micropipette position, could affect this measurerement. The position of the pipette tip is assumed to be fixed during the experiments, but this is not always the case because, due to mechanical relaxation of the microfluidic chamber or of optical components, the real displacement could experience small variations. These movements are usually small compared with the trap displacement, so they are negligible for a single trajectory. However, for repeated trajectories that cycle up and down between two fixed force values, the small displacements could accumulate, and result in a drift of the force-extension curve along the extension axis. It is then necessary to periodically re-align the curves, by applying a shift to the initial curves in such a way to be able to compare the data from different cycles.

The algorithm we developed reads the output files of the mini-tweezers, identifying each distinct trajectory, and stores the information about the time-step, forces, and positions. On the y-axis are gen-

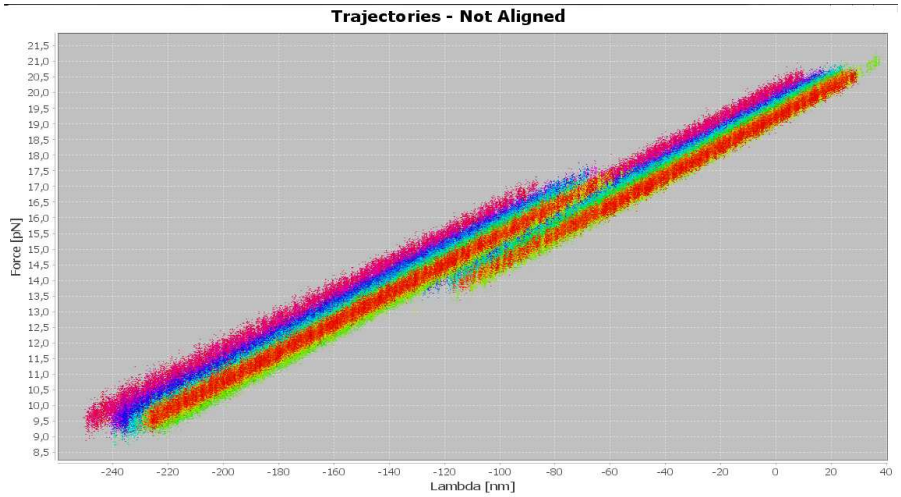


Figure 90: Series of force-extension trajectories during a pulling experiments, before the mathematical alignment.

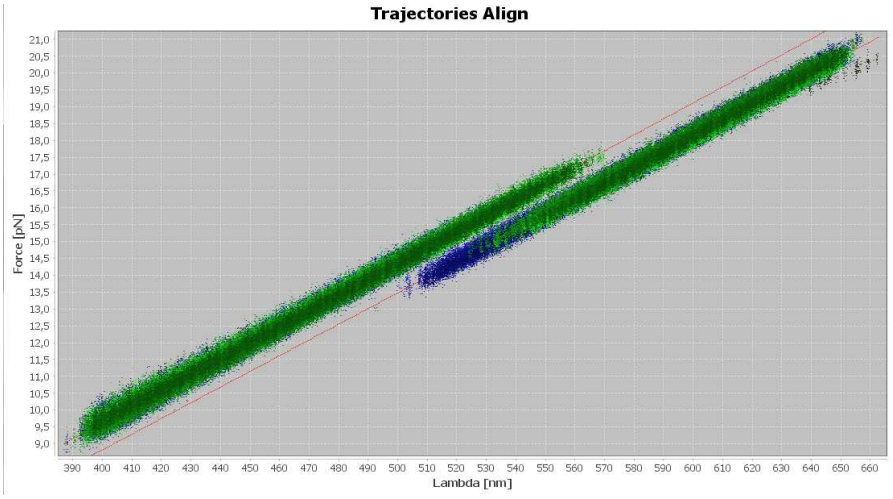


Figure 91: Series of force-extension trajectories during a pulling experiments after the alignment with the two linear fit (red lines) for the folded and unfolded state. In green the unfolding trajectories and in blue the refolding.

erally represented the coordinate along which the trap movements take place. The movements along the other directions are taken to be null, therefore (after regrouping them depending on direction of the displacement (extension or relaxation), is it possible to apply a filter on those trajectories that have orthogonal components greater on average than some threshold value.

Each trajectory then is divided into groups of points, where each group corresponds to different intervals of the force (the same set of intervals for all the trajectories). For each interval, it is measured the average position and from all the trajectories we can extract an "average curve" to be used as reference to align all the others.

Finally, a plot is generated to facilitate the visualization of the force intervals of alignment, in other words the range of forces where the hairpin is for ll trajectories in the folded or unfolded state (Fig.90).

On these force intervals, the program evaluates the average shift as compared to the reference curve. This shift value is applied to align the curves. The new trajectories are plotted (Fig.91) to verify that the chosen intervals produced a correct alignment. Moreover, it is possible to fit the curves in a chosen range, to evaluate the effective stiffness in the folded state and in the unfolded state (Fig.91). Also, a 2D histogram can be produced to verify the probability of each neighbor $[\lambda, \lambda + \Delta\lambda]x[f, f + \Delta f]$; such a plot allows to check the separation between the two states (Fig.92).

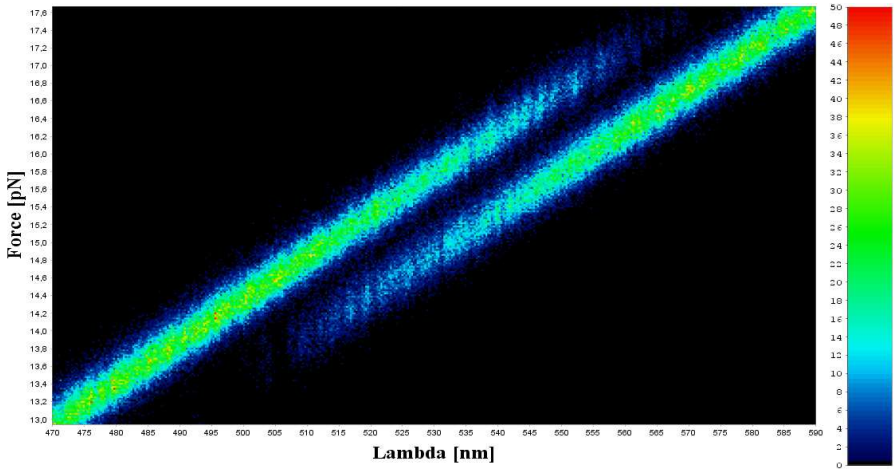


Figure 92: 2D histogram representing the probability to observe a certain combination of force and extension during all the cycles in an experiments.

D

RECOVERING THE TRUE HAIRPIN EXTENSION

The information collected by the optical tweezers, about position and force, refers respectively to the position of the trap, and the force applied on the trapped bead. If we consider the model for the hairpin in the folded state, represented in Figure 18(a), the total extension λ is the sum of the contributions due to all the elements, namely: (i) optical trap, (ii) DNA handles, and (iii) hairpin. This ensemble can be modeled via an effective spring constant given by the formula:

$$\frac{1}{k^{eff}_{fold}} = \frac{1}{k^{b}} + \frac{1}{k^{handles}} + \frac{1}{k^{hairpin}} \tag{D.1}$$

By assuming that the contribution of the hairpin in the folded state to the total extension is described by the orientation of a segment of length equal to the DNA diameter (FJC for a single element, $x^{d}(f)$) the stiffness at a given force f is:

$$\frac{1}{k^{hairpin}_{fold}(f)} = \frac{dx^{d}(f)}{df} = \frac{1}{k_B T}\left[1 - \left(\frac{k_B T}{bf}\right)^{2} - \coth\left(\frac{bf}{k_B T}\right)^{2}\right] \tag{D.2}$$

Now, if the elastic spring constant of the entire system is deduced from the experiment (for example by fitting the trajectories), the additional contributions of the optical trap and of the handles to the total displacement can be estimated.

In the range of forces of the pulling experiments, the soft spring constant of the optical trap dominates the contribution, while the DNA handles are extremely stiff so they contribution to the total variation of length is negligible. Therefore, we could extract the hairpin extension from the formula:

$$\begin{aligned} x_{DNA}(f) + \lambda_0 &= \lambda - \Delta x^{b} - \Delta x^{handles} \\ &\approx \lambda - \frac{\Delta f}{k^{b}} \\ &\approx \lambda - \left(\frac{\Delta f}{k^{eff}_{fold}} - \frac{\Delta f}{k^{d}}\right) \end{aligned} \tag{D.3}$$

where λ_0 is a constant term.

E

NUMERICAL METHODS FOR MOLECULAR DYNAMICS SIMULATIONS

There are three main levels of detail that can lead to quite different molecular dynamics (MD) simulation set up:

- **All-atoms**. In this model each atom is represented as a point particle (in the special case of the *shell model* as two, one for the ion core and one for the electronic cloud). This choice gives the maximum amount of information on a molecule that it is possible to obtain within a classical MD simulation, since each atom contributes in full to the dynamical evolution of the system. On the other side, it also increases the number of degrees of freedom that must be taken into account in the force computation, thus implying larger memory occupation and longer simulation times.

- **United-atoms**. This model attaches the light (and fast) hydrogen atoms to the corresponding cation, thus making up a single unit or pseudo-atom, with specific parameters. Usually, carbons in methyl group ($-CH_3$) and methylene bridge ($-CH_2$) are merged into a single pseudo-atom. This choice suppresses the fast vibrational modes due to the light hydrogens (in many applications, their motion may be irrelevant and only average positions need to be considered), thus reducing the degrees of freedom of the system, and allowing to increase the integration timestep. Obviously this simplification has a cost in terms of detailed description of the molecular evolution and could misrepresent the behaviour of some configurations.

- **Coarse-grained**. This MD model describes the molecules as composed by groups of atoms with different levels of granularity, by replacing entire chemical groups by just a few degrees of freedom. For example, the 10-15 atoms in one nucleobase of DNA may be represented in some model by 3 points and 2 angles. Such techniques have found considerable application in computational biophysics, because they can very significantly increase the system size, allowing the exploration of much longer timescales and phenomena otherwise inaccessible to conventional MD simulation. It is also for this kind of approach that M. Levitt, A. Warshel and M. Karplus were awarded the 2013 Nobel Prize in Chemistry, notably *"for the development of multiscale models for complex chemical systems"*.

Once the degrees of freedom (that is, level of detail) and the equations of motion are defined, the main steps common to any MD simulation are :

1. definition of an *initial configuation* of the system, where actual of pseudo atoms are described as point-like objects;
2. choice of the *force field* describing the empirical interatomic force laws (both functional form of the potential and the specific parametrization), to reproduce at best the real system properties of interest;
3. start a time-integration loop with a fine time-step (1 to a few 10^{-15} s), in which at each step:
   a) calculate the *force* acting on each particle $f_i = -\partial_i V_{eff} + R_i + F_i$, including possible contributions R_i from the numerical thermostat and/or barostat, and eventual external perturbing forces F_i;
   b) update the *position* and *velocity* of each and every particle to the next time-step, according to a discrete-time integration algorithm;
   c) accumulate estimates for the *physical observables* during the course of the phase-space trajectory.

E.1 INITIAL CONFIGURATION AND THE PERIODIC BOUNDARY CONDITIONS

The choice of the initial configuration of a molecular structure is done by fixing the position and velocity of all the atoms in the system. Careful attention must be posed in this choice to avoid unrealistic configurations, therefore a energy-minimization procedure is the first step required before any MD run. At the macroscopic scale the surface effects due to the limited box size are normally considered negligible, but this is not true at the nanometer scale. To avoid surface effects it is possible to introduce *periodic boundary conditions* (PBC), which make the system virtually of infinite extension.

The periodic-boundary conditions assume that the system is simply homogeneous for an infinite length scale, meaning that the system is not truly isolated (in the thermodynamic sense) but is the unit cell of a lattice of infinite copies of the same box, periodically repeated along one, two or three spatial directions. The practical effects are that a particle that escapes from one side of the periodic box will emerge from the opposite side, with the same velocity. Moreover, particles close to the periodic borders could also interact with virtual copies of the other atoms. This method requires great care, to avoid unphysical interactions between an atom and its own replica: the box size,L, must be chosen at least twice longer than the cut-off radius of the most extended interaction potential (see E.3) $L \geqslant 2r_{cutoff}$.

The DNA is a complex biological molecule composed by subunits, the nucleotides, each one with its own internal structure. To represent the molecule in a all-atoms MD simulation it is then necessary to describe the three-dimensional position of each atom in the molecule, and then to assign interaction relations that correctly reproduce the molecule structure and behavior. The configuration generated by the

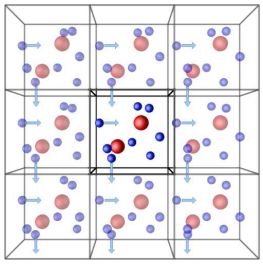


Figure 93: Schematic of the periodic boundaries conditions for a 2D square cell. In this case, the primary box is replicated on a plane.

software is a ProteinDataBank[1] file. By using specific software it is then possible to modify the initial structure, create defects, deform the shape of the DNA strand, add the water and ions,... in such a way to obtain an initial configuration with the desired properties.

E.2 FORCE FIELDS

The force field are the files that describe the empirical interatomic force laws, to reproduce at best the real system properties. According to their physical meaning, the interaction can be divided into different categories:

- *Bonded interactions*, describing the covalent bonds between atoms in a molecule. Typically, these interactions are decomposed into 2-body terms (depending on the distance between pairs of atoms), 3-body terms (depending on the angle formed by triplets of atoms) and 4-body (acting on the dihedral angles formed by clusters of four atoms). Bonding interaction are predefined for a fixed number of degrees of freedom for each atom and remain fixed (unbreakable bonds) during the entire simulation.

- *Non-bonded interactions*, such as long-range Coulomb or Van der Waals forces. These are simpler pair interactions, but they require particular attention in the numerical and algorithmic implementation since they are in principle extended to all the atoms in the system. This could lead to very time-consuming algorithms, poorly scaling with the system size (at worst $O(N^2)$). Nevertheless, by applying specific techniques (see E.3), it is possible to reduce the computational time necessary to evaluate these contributions, possibly down to $O(N)$ algorithms.

- *Constraints* and *external forces*. These can be imposed to the whole system and give an external contribution to the total energy and are null for unconstrained isolated system. Nevertheless they could be useful to simulate the interaction with thermostats, barostats or in presence of external force fields, as in the case of steered-MD that will be used to simulate optical-tweezer pulling experiments.

1 Protein Data Bank (pdb) file format is a textual file format describing the three-dimensional structures of molecules, as protein and nucleic acid, held in the Protein Data Bank.

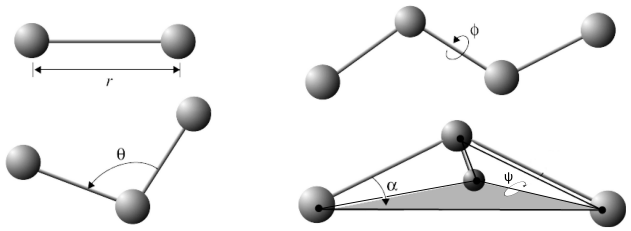


Figure 94: Bonding interactions. Schematic representation of the pair, bond-angle, proper dihedral-angle, and improper-dihedral angle interactions, usually described by harmonic or cosine potential functions.

Bonded interactions play a central role in defining the secondary structure of biomolecules like DNA. These forces are described by a set of (usually) simple potential functions (Figure 94), to assign two-particle bond length, three-particle preferred angle, four-particle dihedral angle, and calibrate their relative strength. The latter are often expressed in the form of simple harmonic or periodic (sine/cosine) functions, in some cases adding higher-order (cubic, etc.) corrective terms. Additional terms can be introduced to stabilize particular polyatomic shapes, such as pentagonal or hexagonal aromatic cycles.

Non-bonded interaction forces, which are largely responsible for the tertiary structure, are described in more complex terms, demanding special care in the calculation of the resulting forces and in the setting of the time integration algorithms, and are described in some detail in the Sections E.3,E.4 E.5.

Any molecular dynamics software then introduces the appropriate numerical techniques to reduce the computational complexity of short-range non-bonded interaction calculations in the system (such as Van der Waals forces), or the short-range part of the electrostatic interaction (when using Ewald summation).

E.3 THE INTERACTION CUT-OFF

Non-bonded forces are in principle extended to all particles in the system, with a functional shape of the corresponding empirical potential that becomes zero only at infinity. However, the rapidly decreasing potential function (dispersion forces decay as $1/r^m$ with $m \geqslant 6$; the $1/r$ of pure Coulomb is practically screened by surrounding charges, down to a $\sim \exp(-ur)/r$, as accounted in the Ewald-sum technique) make the numerical error comparable to the force values after a relative short range, so that the average effect of "distant" particles becomes negligible. It is therefore possible to introduce a cut-off distance, r_c, beyond which the potential functions are practicaly truncated to zero. The value of r_c must be such as to avoid introducing unphysical effects, such as the impulsive force (infinite derivative) at the cut-off. In the numerical implementation of the algorithm, smoothing techniques can be applied to avoid this issue, for example modifying

the potential tail near r_c by a $C(2)$ or higher-order function, so that the force remains small and continuous up to r_c.

The introduction of a force cut-off reduces the number of neighbors to take into account for each particle, to those within its reduced "sphere of interaction" (Figure 95); this is a small coordination number $z << N$, which lowers the complexity of the search algorithm to $O(zN)$.

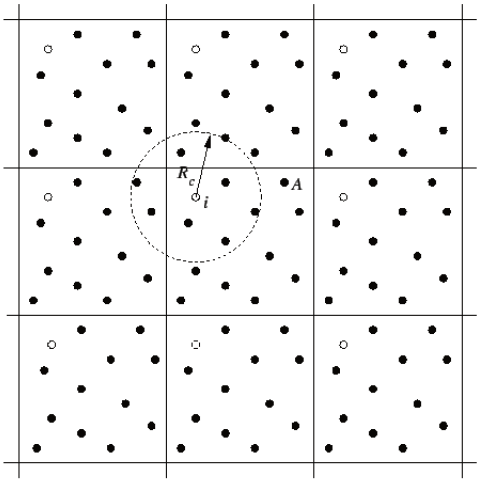


Figure 95: Sphere of interaction for the particle i (diameter= $2r_c$); with the periodic-boundary condition, the sphere includes both "real" particles, in the central box, and "image" particles, from the periodically-repeated surrounding boxes.

E.4 VERLET'S LISTS AND LINKED-CELLS

In principle, neighbor search for computing non-bonded forces requires a $O(N^2)$ algorithm, since all pairs of atoms must be tested, to include them in the energy if their relative distance falls within the force cut-off. Various schemes have been devised to reduce the complexity of this brute-force pair search, which can become overwhelming with increasing N; even after introducing a force cut-off, the code does not know a priori how many and which particles fall in the interaction sphere. In the *Verlet's list* method, are recorded the labels of all atom pairs that satisfy the condition:

$$List_{i-atoms} = \{j \in 1, 2, ...N \mid |\vec{r}_i - \vec{r}_j| < r_c + r_b\} \tag{E.1}$$

The new quantity r_b is a "buffer" distance (Figure 96) to prevent particles from entering or exiting abruptly in the interaction sphere

$$|\vec{r}_i - \vec{r}_j| < r_c$$

thereby causing sudden jumps in the energy and impulsive force contributions, whenever one particle suddenly gets in or out the sphere.

An integer array (list) is kept for each particle $List_{i-atoms}$, which is periodically updated by considering that before a new particle enters the outer buffer and makes it to the inner sphere, several Δt time

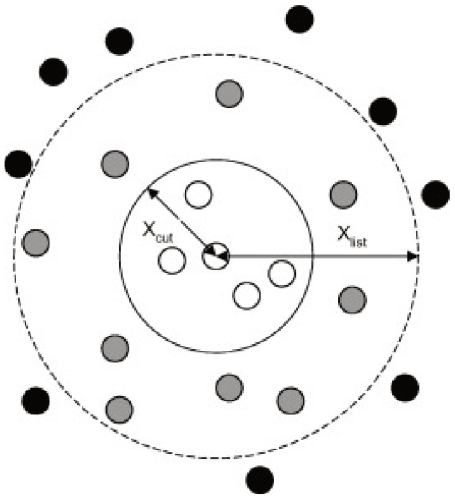


Figure 96: Schematic representation of the Verlet's lists: in white, particles having a non-negligible interaction with the central one; in gray, particles included in the list but with negligible interaction; in black, particles not included on the list.

steps can intervene. The refresh of the list is done when the maximum distance covered by a particle is greater than half of the buffer distance.

The updating of Verlet's lists is an algorithm of order $O(N^2)$, so it would seem obvious that to reduce the total computation time it is better to increase the buffer distance r_b. However, the number of extra particles included in the list increases rapidly with r_b so that a too large r_b would frustrate the benefits from the introduction of a relatively short r_c.

A further optimization is obtained with the *cell linked-list* algorithm. This technique consists in subdividing the simulation box into cells with an edge length r_{lc}, greater than or equal to the Verlet radius. This reduces the number of particles to be checked during the updating of the Verlet lists: in fact, it is possible to consider just the particles that belong to the same cell of the central particle, plus those in the neighboring cells (8 cells in two dimensions, or 26 in three dimensions). This technique is useful only if the initial system could be subdivided in more than 27 cells (the number of repeated images plus the central box, 3x3x3 in three dimensions), that is if the system size is larger than $6r_c$ in all three directions. The cell lists are updated by the same Verlet algorithm, but both operations are done only on the neighbor cells of each cell, thus effectively reducing the order of the algorithm to $O(N)$.

E.5 INTEGRATING THE EQUATIONS OF MOTION

The equations of motion are second-order differential equations. It is then possible to use the numerical methods developed for solving the ordinary differential equations, to find numerical approximations to the dynamical evolution of the atomic system.

Once the force field is chosen, and the set of equations representing the external and internal constraints eventually imposed on the system are defined, it is possible to compute the total force acting on each and every atom due to its interaction with other atoms of the system, as:

$$\vec{F}_i = -\vec{\nabla}_{r_i} V_{eff}(\vec{r}_1, ..., \vec{r}_N) \tag{E.2}$$

the gradient being taken with respect to the coordinates of the atom i. Because of the simultaneous dependence of V_{eff} on the coordinates of all atoms (even if this dependence can be systematically reduced to 2-, 3-, and 4-body pair interactions), the equations of motion (4.4) are a strongly coupled set of partial differential equations. The corresponding system does not have an analytical solution, making it necessary to use numerical techniques to evaluate the system trajectory. "Forward" (in time) integration algorithms discretize the time into intervals Δt, typically of the order of 1 fs, and evaluate trajectories by moving the particles step by step, in accordance with the discretized equations of motion.

There are different possible numerical algorithms to perform the numerical integration, such as the *velocity-Verlet*, the *Euler*, or the *leap-frog*. The general idea is to use a Taylor series expansion for the initial partial differential equation with respect to time. Consider a small time interval between any two subsequent time steps; by using the analogy with the infinitesimal derivatives, $dx(t)/dt = v(t)$ and $d^2x(t)/dt^2 = a(t) = f(t)/m$, the position of a particle at time $t + \Delta t$ is:

$$x(t + \Delta t) = x(t) + \Delta t\, x'(t) + \frac{\Delta t^2}{2} x''(t) + O(\Delta^3) \tag{E.3}$$

allowing to approximate the new position from the finite-difference increments of position x, velocity v, and force f at the previous time step. For the sake of practical implementation, it is worth noting that by combining the forward and backward finite-difference, it is possible render the equation not dependent on the velocities, so that only forces and positions appear in the numerical algorithm (more details are given in the Appendix E).

Such type of algorithms are the core of any MD simulation code. Three of the most widely used ones are:

- *Verlet algorithm*, it only generates positions, and velocities are not needed to compute the trajectory.

  $$x(t + \Delta t) = 2\, x(t) - x(t - \Delta t) + \Delta t^2\, a_i(t) + O(\Delta t^4). \tag{E.4}$$

  Anyway, knowing the particle velocities is useful for estimating the kinetic energy, and other related properties of the system. It is then possible to compute velocities from the updated position,as:

  $$v(t) = \frac{x(t + \Delta t) - x(t - \Delta t)}{2\Delta t} + O(\Delta t^3) \tag{E.5}$$

even if this slightly reduces the efficiency of the algorithm, since it requires to store three position arrays for each particle, $x(t+\Delta t)$, $x(t)$, $x(t-\Delta t)$

- *Velocity-Verlet Algorithm*, it describes the update of the positions and velocities simultaneously:

$$x(t+\Delta t) = x(t) + \Delta t\, v(t) + \frac{\Delta t^2}{2} a(t) + O(\Delta t^3) \tag{E.6}$$

$$v(t+\Delta t) = v(t) + \Delta t \frac{(a(t) + a(t+\Delta t))}{2} + O(\Delta t^3). \tag{E.7}$$

Its two key properties are the *time-reversibility* (upon reversing the direction of integration the same trajectory is obtained) and the *symplectic* nature (conserve by construction the total energy of the dynamical system).

- *Leapfrog algorithm*, it is similar to the Velocity-Verlet and maintains the same properties (time-reversibility and symplectic nature), but it evolves position and velocities at different time-steps:

$$x(t+\Delta t) = x(t) + \Delta t\, v(t+\Delta t/2) \tag{E.8}$$

$$v(t+\Delta t/2) = v(t-\Delta t/2) + \Delta t a(t) + O(\Delta t^3). \tag{E.9}$$

E.6 LANGEVIN DYNAMICS AND TEMPERATURE/PRESSURE ALGORITHMS

In a non-isolated system it is possible to introduce extra terms to the Newton's equations of motion, to simulate the interaction of our system with an external field or with a thermal bath (heat exchange with a fictitious thermostat). The *Langevin dynamics* theoretical construction is extensively used, both in its theoretical formulation, as we saw in the discussion on the transition-state theory, or in numerical simulations, to mimic the effects of a viscous solvent, or of heat exchange with a thermostat. The Newton's equations of motion are modified by adding, for each particle, a viscous background and a fluctuating force describing the thermal bath:

$$\begin{cases} \frac{d\vec{r}_i}{dt}(t) = & \vec{v}_i(t) \\ \frac{d\vec{v}_i}{dt}(t) = & \frac{1}{m_i}\vec{f}_i(t) - \gamma m_i \vec{v}_i(t) + \vec{\eta}_i(t) \end{cases} \tag{E.10}$$

where γ is the parameter that describes the friction due to the viscous background (Stokes's law), and $\vec{\eta}(t)$ the force generated by the thermal bath, that must satisfy the *fluctuation-dissipation theorem*:

- $< \vec{\eta}_i(t_1) >= 0$.
- $< \vec{\eta}_i(t_1)\vec{\eta}_j(t_2) >= 2\gamma m_i T \delta_{ij}\delta(t_1 - t_2)$.

The effect of this algorithm is to slow down by friction the "hot" particles with kinetic energy higher than the fixed temperature T,

and accelerate via a white noise the "cold" particles with kinetic energy lower than T. By using this equation it is possible to generate a "Langevin thermostat", and the system satisfies the requirement of Markovian dynamics.

Other versions of the thermostat (to control the temperature) and barostat (to control the pressure) could be implemented, by modifying the equations of motion with appropriate extra terms, or by periodically rescaling coordinates and/or velocities to the target values. The most commonly used numerical thermostats are:

- **Velocity Scaling algorithm**: all the velocities are scaled by the factor $\lambda(t) = \sqrt{\frac{T^{target}}{T^{inst}(t)}}$ in such a way that the instantaneous temperature T^{inst} always equals the target temperature T^{target}:

$$v(t) \leftarrow \lambda(t)v(t). \tag{E.11}$$

  One obvious limitation of this simple method is that the generated trajectories do not reproduce correctly the constant-{NVT} ensemble, because no fluctuations in temperature are allowed.

- **Berendsen's thermostat**: velocities are scaled at each step, such that the rate of temperature change is proportional to the difference between the instantaneous temperature and the target temperature, via a coupling parameter τ:

$$\frac{dT^{inst}}{dt} = \frac{1}{\tau}\left(T^{target} - T^{inst}(t)\right) \tag{E.12}$$

  Notice that for $\tau \to \infty$ the microcanonical ensemble (constant-{NVE}) is recovered, while for $\tau \to 0$ unrealistically low temperature fluctuations are introduced.

  The change in temperature at each time step is equal to:

$$\Delta T(t) = \frac{\Delta t}{\tau}\left(T^{target} - T^{inst}(t)\right) \tag{E.13}$$

$$= \left[\frac{\Delta t}{\tau}\left(\frac{T^{target}}{T^{inst}} - 1\right) + 1 - 1\right] T^{inst}(t). \tag{E.14}$$

  Besides, we also have:

$$\Delta T(t) = \sum_i \frac{m_i \lambda(t)^2 v_i^2}{3Nk_B} - \sum_i \frac{m_i v_i^2}{3Nk_B} \tag{E.15}$$

$$= \left(\lambda(t)^2 - 1\right) T^{inst}(t). \tag{E.16}$$

  Thus, the scaling factor λ for the particle velocities can be determined as:

$$\lambda(t)^2 = \frac{\Delta t}{\tau}\left(\frac{T^{target}}{T^{inst}(t)} - 1\right) + 1. \tag{E.17}$$