SINGLE-MOLECULE STUDIES OF NUCLEIC ACIDS
AND THEIR PROTEINS
Single-Molecule Studies of Nucleic Acids and Their Proteins

David Bensimon
*LPS-ENS, PSL, CNRS, France and Department of Chemistry and Biochemistry, UCLA, USA*

Vincent Croquette and Jean-François Allemand
*LPS-ENS, PSL, CNRS, France*

Xavier Michalet
*Department of Chemistry and Biochemistry, UCLA, USA*

Terence Strick
*IBENS, PSL, CNRS, France*
This work is dedicated to our spouses for their patience and support.
Preface

Biophysics and biochemistry are undergoing a revolution with the development of new tools and techniques to manipulate and observe single molecules and their interactions. In this monograph, our purpose is not to present a comprehensive review of this vast and rapidly expanding field, but rather to introduce these novel techniques and exemplify their use in the study of nucleic acids and their interactions with proteins. In other words, our goal is to provide an introduction to the main techniques which are used to study single molecules, both via their manipulation and visualization, to allow the reader to understand the concepts behind these methods.

We discuss in detail the mechanical properties of nucleic acids, where a comprehensive understanding was achieved thanks to the use of single-molecule manipulation approaches. This agreement between theoretical models and observations has set the stage for the single-molecule study of the interaction between nucleic acids (DNA or RNA) and their associated proteins.

While we will not review this quickly expanding area, we will exemplify the use of single-molecule manipulation and visualization methods in the study of some specific DNA motors: DNA and RNA polymerases, helicases, and topoisomerases. Our purpose is to show that these techniques can open up new vistas on the understanding of the mechanism of molecular motors and be used to address a variety of questions pertaining to their rate, processivity, step size, efficiency, stoichiometry, etc.

We acknowledge that in the choice of the examples presented here, we have been biased by our own investigations and apologize to our colleagues for perhaps not giving proper exposure to their work. We believe that a comprehensive review would have been impossible at that stage.
Acknowledgements

The work presented here would have been impossible without the help of our many friends, students, post-docs, and colleagues, and in particular: L. Jullien, S. Weiss, F. Ding, M.N. Dessinges, T. Lionnet, C. Gosse, G. Lia, E. Praly, F. Mosconi, A. Meglio, E. Cavatore, M. Peliti, B. Maier, O. Saleh, K. Neuman, N. Dekker, M. Manosas, R. Lavery, T. Owen-Hughes, L. Finzi, M. Duguet, N. Cozzarelli, N. Crisona, R. Ebright, A. Volgodskii, S. Benkovic, F.X. Barre, and E. LeCam. We thank C. Bagshaw for a critical reading of this manuscript and making many suggestions. We are also grateful for the support of our institutions: CNRS, PSL, ENS, Univ. Paris 6 and 7, UCLA, and our granting agencies: ANR, PUF, INSERM, EU (via FET-Open, ERC, Marie-Curie, and EURYI grants), HFSP, NSF and NIH.
## Contents

5.3 Unzipping DNA 113  
5.4 Conclusions and perspectives 116

### 6 Single-Molecule Studies of DNA-Associated Motors 119

6.1 Introduction to DNA–protein interactions 119

### 7 DNA and RNA Polymerases 121

7.1 Introduction to polymerases 121  
7.2 An introduction to replication 121  
7.3 Single-molecule *in vitro* DNAP experiments 123  
7.4 *In vivo* studies of single replication complexes 132  
7.5 An introduction to transcription 133  
7.6 Single-molecule *in vitro* RNAP experiments 139  
7.7 *In vivo* studies of transcription 151  
7.8 Conclusions and perspectives 153

### 8 DNA Helicases 155

8.1 Introduction to helicases 155  
8.2 Bulk versus single-molecule assays: the case of RecBCD 158  
8.3 Active and passive unwinding mechanisms: RecQ versus gp41 160  
8.4 The UvrD helicase studied by single-molecule manipulations 164  
8.5 Single-molecule FRET studies of the Rep helicase 171  
8.6 Problems 174

### 9 Topoisomerases 177

9.1 Introduction to topoisomerases 177  
9.2 Bulk assays 180  
9.3 Single-molecule *in vitro* assays: type I topoisomerases 181  
9.4 Single-molecule *in vitro* assays: type II topoisomerases 185  
9.5 Single-molecule study of the coiling activity of gyrases 195  
9.6 *In vivo* studies of single topoIV localization and action in *E.coli* 197

### 10 Conclusions 199

*References* 201  
*Index* 223
Introduction to DNA

1.1 Introduction

When, in 1953, Watson and Crick suggested that deoxyribonucleic acid (DNA) adopted a double-helical structure [1], the molecule was already known to be the support of genetic heredity [2, 3]. However, the discovery that DNA consisted of two complementary polymers (two strands) winding around each other significantly changed the way that we thought about cellular processes, such as DNA replication [4]. It provided a much needed molecular and structural basis for DNA replication and transcription. It became clearer that the study of molecular interactions was a necessary step in understanding their cellular function. In the thirty years following the discovery of the double helix, numerous techniques have emerged to advance this study, which by now constitute the bulk of “Molecular Biology”. These techniques allow for the transforming, synthesizing, and sequencing of DNA molecules as well as for studying and quantifying the interactions between biomolecules (such as proteins and DNA). A landmark in this genomic era was set when the human genome sequencing programme reached its goal less than fifty years after Watson and Crick’s discovery.

The decade beginning around 1990 has witnessed the emergence of a wealth of new techniques and tools devoted to the physical study of single molecules. Methods as diverse as optical, acoustic, and magnetic tweezers, microfibres, and atomic force microscopy are used in many labs to manipulate (move, stretch, or twist) single biomolecules (DNA, proteins, carbohydrates, etc.). In parallel, optical methods based on fluorescence (by resonant energy transfer or directly with evanescent waves, two photons, or confocal configurations) have also been developed to study biochemical processes at the single-molecule level. Many groups are combining both aspects to monitor or affect the activity and visualize the displacement of a single molecule (myosin, RNA polymerase, etc.) in vitro and in vivo.

From a physicist’s point of view, such experiments have provided new data against which to test models of polymer elasticity, with or without specific types of interactions, such as electrostatic, self-avoidance, or base-pairing. Single-molecule studies have also provided quantitative constraints on more complex polymer problems, such as protein folding.
The mechanical properties of nucleic acids and proteins play a fundamental role in their biological function. For instance, the double-helical structure of DNA stores the genetic content of the molecule within the core of the double helix, thus preventing proteins easy access to the genetic code. It is, however, clear that the initial steps of many fundamental biological processes (such as DNA replication and the transcription of DNA into messenger RNA) depend on the unwinding, or melting, of regulatory DNA sequences. Once the DNA has been opened locally by a combination of mechanical and enzymatic effects, the proteins can then read and copy the genetic code. As we shall see in the following, the use of single-molecule approaches has made it possible to quantitatively reproduce and analyse such effects, leading to a much better understanding of the role these structural transitions play in vivo.

1.2 Single-molecule versus bulk studies

This book is devoted to a review of the interactions between some proteins and DNA at the single-molecule level. Since DNA/protein interactions have been extensively studied for the past fifty years using ensemble (bulk) enzymology methods, one may legitimately wonder what can be learned from single-molecule investigations. The short answer is individual behaviour. In ensemble assays, the measured observable is an average over a population of enzymes that may be extremely heterogenous. Some enzymes may be inactive, others might not be bound to their substrate, while those that are bound will usually be in different stages of their catalytic cycle, such as a pause state. As a consequence, the enzymatic activity deduced from these studies often significantly underestimates the “real” activity of a single enzyme.

Moreover, since enzymes are stochastic machines (their passage from one state to the next is a probabilistic process), an ensemble of enzymes cannot be synchronized over many catalytic cycles. This fact sets a fundamental limit to a detailed study of enzymatic mechanisms at the ensemble level. However, these details are in principle accessible when investigating the activity of a single enzyme. For example, the step size of a translocating enzyme can be directly observed in single-molecule but not ensemble studies. Single-molecule assays allow the study of subpopulations in an ensemble of enzymes. In fact, by providing the whole probability distribution of enzymatic activities, single-molecule assays provide in principle all the information that can be obtained on these molecular machines.

The great difficulty of single-molecule assays is putting these considerations into practice. Practical issues (temporal and spatial resolution, Brownian fluctuations, noise, photobleaching of dyes, etc.) set other limits on these studies. Moreover, the degree of parallelism of single-molecule assays is, at present, very limited (molecules are usually studied one at a time), so the sampled population is quite small (at best, a few hundred molecules). As a result, the statistical uncertainties on the measured observables are quite large (> 5%), and certainly much larger than what can be measured in ensemble studies. It is, for example, very tedious to extract basic enzymology parameters (such as the
Michaelis–Menten parameters of a first order enzymatic reaction) from single-molecule assays. Similarly, ATP consumption rates are much easier to measure in bulk studies than in single-molecule assays.

One must, therefore, consider single-molecule assays as a useful addition to the toolbox of enzymologists. They can provide details on the enzymatic mechanism which are impossible to get using standard approaches, but yield much less precise average kinetic parameters than the latter (although they may be more accurate if the bulk study is dominated by a large subpopulation of inactive enzymes).

1.3 DNA structures

Most biomolecules (DNA, RNA, proteins) are linear polymers—i.e., they consist of a linear chain made of repeating structural subunits. For DNA and RNA, the monomeric subunits are nucleic acids of four different types (adenine, guanine, cytosine, and thymine (uracil for RNA)). For proteins, the building blocks are the twenty different amino acids. The structure of the biomolecule is an essential determinant of its function. This is particularly true for proteins for which the prediction from the sequence of the structure (and probable function) of the molecule has emerged as a major goal of the post-genomic area. This is also true for RNA whose 3D structure plays a crucial catalytic role in many reactions, such as the synthesis of proteins by the ribosomal complex.

In contrast with these examples in which the structure of the biomolecule is apparently uniquely determined by the sequence of its subunits, the structure of DNA is a double helix independent of its sequence. This structure has profound implications on all genetic processes. For transcription to occur, the two strands of DNA have to be locally unwound by an RNA polymerase, a process that might sensitively depend upon the torque acting on the molecule. During DNA replication, the two strands have to be unzipped by a helicase to make way for the enzymatic complex polymerizing two new DNA strands on the templates provided by the unwound single strands. This DNA unwinding generates a large torque ahead of the complex which is released by specialized enzymes, the topoisomerases. These enzymes are also responsible for disentangling DNA prior to cell division. Understanding how these enzymes and others work, therefore, requires an understanding of the mechanical properties of DNA.

Before studying the elastic behaviour of DNA in greater detail, it is important to recall its physico-chemical properties. DNA can be described as a molecule having a hierarchical series of structures. The primary structure describes the chemical/genetic sequence of the molecule. The secondary structure describes the way that the two constituent strands of the molecule wrap around each other to form the double helix. The tertiary structure represents the trajectory of the double helix’s axis: the molecule might, for example, writhe about itself or wrap around globular histone proteins to form the chromatin fibres which are the main constituents of chromosomes. These higher-level structures are less well known, and would correspond to the series of packaging steps by which DNA is organized into a chromosome.
1.3.1 **Primary structure of DNA**

In its most common B-form, DNA consists of two anti-parallel strands wrapping around each other with a right-handed helical symmetry. Each strand is made of a polymer chain of deoxyribose rings linked by a phosphodiester bond between their 3′-OH moiety and the phosphate group on the 5′ carbon of their neighbour (see Fig. 1.1a). The distance along a strand between nearby phosphate groups is about 7 Å. One of four possible different chemical groups (bases), adenine (A), guanine (G), cytosine (C), and thymine (T) is linked to the 1′ carbon of the sugar ring. These groups consist of planar and hydrophobic aromatic cycles, resulting in a strong hydrophobic attraction between neighbouring bases. In an analogy of the piling of plates, this attraction is referred to as the “stacking” interaction.

1.3.1.1 **Secondary structure**

The presence of primary and secondary amines on the aromatic cycles of the bases enables guanine (G) and cytosine (C) to form three hydrogen bonds and adenine (A) and thymine (T) to share two hydrogen bonds (see Fig. 1.1a). These bases are complementary: the adenine (resp. guanine) group on one strand is paired with a thymine (resp. cytosine) group on the adjacent (complementary) strand. Hydrogen bonding between A (resp. G) and T (resp. C) yields a stable right-handed helix of complementary and anti-parallel strands (see Fig. 1.1b), which, from an information point of view, are to each other as a film is to its negative. These pairing and stacking interactions contribute to the exceptionally high bending rigidity of double-stranded DNA as compared to artificial polymers, such as polyethylene (see section 4.1).

The global structure (handedness, helical pitch, effective diameter, etc.) adopted by the double helix depends on a number of environmental parameters, such as solvent, ionic conditions, and temperature [6]. The canonical form of DNA in solution is named B-DNA and exists in the right-handed double helix just described. B-DNA has a diameter of about 24 Å, vertical spacing between bases of roughly 3.4 Å, and helical pitch of 10.4 base pairs (bp) per turn. The various interactions which stabilize the B-form of DNA also give it local rigidity: its persistence length $\xi$ (i.e., the distance over which the molecule is more or less straight in spite of thermal agitation that tends to shake and bend it, see section 4.1) is about $\xi = 50$ nm or 160 bp. This is much larger than the persistence length of most synthetic polymers, such as polyethylene, polystyrene, PVC, etc., which have a $\xi \sim 1$ nm; dsDNA is also much stiffer than other natural polymers, such as RNA or single-stranded DNA (ssDNA). An RNA or ssDNA molecule can adopt complex shapes and change direction over distances of a few bases, forming, for example, “hairpin” structures with a double-stranded stem ending with a single-stranded loop.

Let us mention a few non-canonical forms which may appear punctually within B-DNA:

- If the bases on the two opposite strands are not complementary, a “bulge” of unpaired bases may appear.
Fig. 1.1 Primary, secondary, and tertiary structure of the DNA molecule. (a) Each of the two strands of the DNA molecule is made of a ribose–phosphate backbone. Chemical groups (adenine (A), guanine (G), cytosine (C), and thymine (T)) are attached on the ribose sugar. The distance between base pairs is 3.4 Angstrom for the standard B-form of DNA. Note that the two strands are anti-parallel: orientation can be defined using the anchoring point of the phosphate group (carbon 5′) on the sugar. (b) Secondary structure of DNA: each strand wraps around each other in a non-symmetric helical fashion. The pitch of the helix is 10.5 base pairs for the standard B-DNA, which means that the twist angle is \( \approx 36^\circ \). (c) At a larger scale, DNA behaves as a worm-like chain whose orientation decorrelates over a typical lengthscale of \( \xi = 50 \) nm. (d) Long DNA molecules (\( L \gg \xi \)) look like random coils with a typical gyration radius \( R_g = \sqrt{2\pi L} \approx 10 \) \( \mu m \) in the case of an E.coli chromosome (four million base pairs).
• In the presence of a palindromic sequence, the DNA may locally adopt a “cruci-form” shape.
• A denaturation bubble is formed if the base pairs facing each other on the two strands of the double helix are complementary but dissociated.

X-ray diffraction data collected on DNA crystallized in various conditions of humidity or salinity show that the double helix is structurally polymorphic. For example, if the molecule undergoes crystallization in low humidity and in the presence of low amounts of cations, a structure known as A-DNA is observed. It is more compact than the B-form (the spacing between bases is 2.6 Å [6]), but has the same handedness. There also exists a left-handed double helix with a pitch of twelve base pairs per turn [6], known as Z-DNA. This DNA structure not only requires strong ionic conditions to be stabilized, but also a sequence of alternating purines (i.e., adenine or guanine) and pyrimidines (i.e., thymine or cytosine).

The fact that the bases in B-DNA are buried inside the double helix confers a very low reactivity to the molecule and drastically reduces the interactions of the polymer with itself (in contrast with single-stranded DNA or RNA which can form complex tertiary structures, such as hairpins). This feature is essential to preserve the genetic code from being damaged or modified easily, but it presents a challenge to the protein machinery which has to regulate, edit, copy, and transcribe the molecule. This machinery must access the bases and, therefore, open the DNA molecule.

As originally mentioned by Watson and Crick [1] and proved by Meselson and Stahl [7], the existence of two copies of the genetic information allows for a simple DNA replication mechanism (the so-called “semi-conservative” replication model of DNA): each original strand is used as a template for the synthesis of its complementary strand. Another essential feature of the two strands’ complementarity is to allow DNA mismatches and mispairing to be corrected using the original strand as a template. This ensures the extremely high fidelity of DNA replication (the error rate is of the order of one base per $10^9$).

The DNA replication mechanism raises a major question: if each strand is used as a template for a new molecule, the mother DNA has to be unwound to be duplicated but then how are the daughter strands separated “without everything getting tangled” [4]? This problem was solved by the discovery of enzymes, known as topoisomerases, which are capable of unknotting and untangling the DNA molecule. These topoisomerases control both the torsional stress in DNA and its overall topology.

### 1.3.2 Tertiary structures in DNA

The primary and secondary structures of DNA confer upon the molecule a large bending stiffness, so that the typical lengthscale, denoted by the persistence length $\xi$, over which thermal fluctuations are able to bend the axis of the molecule under physiological conditions, is about 50 nm (as compared to 1 nm for usual polymers). In other words, if $\vec{r}(s)$ and $\vec{r}(s')$ are two vectors tangent to the DNA’s axis at positions $s$ and $s'$, $\xi$ is the typical decorrelation length (see Fig. 1.1c):
\( \langle \vec{t}(s) \cdot \vec{t}(s') \rangle = e^{-|s-s'|/\xi} \)

DNA can thus be modelled as a flexible rope whose random coil geometry constitutes de facto the tertiary dynamical structure of DNA. Given the crystallographic length \( L \) of the DNA, one can estimate the typical radius of gyration \( R_g \) of the molecule using the model of a “semi-flexible” polymer in solution [5], see section 4.1:

\[
R_g = \sqrt{2\xi L}
\]

For the \textit{E.coli} chromosome, \( L \approx 1.5 \text{ mm} \) so \( R_g \approx 10 \mu\text{m} \) (to be compared with the typical size of the bacteria \( \approx 1 \mu\text{m} \)), see Figs 1.1d and 1.2.

Therefore, to fit into the cell, DNA has to be packed. The way this packaging is achieved implies major changes in the DNA topology. In bacteria, the space occupied by the circular chromosomal DNA is minimized by maintaining the molecule in a highly supercoiled state, exhibiting compact braided structures known as “plectonemes”, see Fig. 1.2, reminiscent of the interwindings observed in a tangled phone cord.

In eukaryotes (organisms possessing a nucleus), DNA is wrapped about protein octamer complexes, known as histones, with about 145 bp of DNA wrapped around each histone to form a nucleosome core. Such nucleosomes occur every 200 bp along the DNA to form the bead on a string structure of chromatine. Chromatine itself is further compactified into supercoiled structures of higher order, eventually resulting in the packaging of a total DNA about 1 metre long (in humans) into a cell nucleus a few microns in diameter, see Fig. 1.3. One of the major challenges of biology is to understand how this extremely compacted molecule is regulated, copied, transcribed, repaired, and disentangled prior to cell division.

\[\textbf{Fig. 1.2} \quad \textit{E.coli} and human chromosomal DNA. (a) Sketch of a supercoiled \textit{E.coli} chromosome: its total length is about 1.5 mm. (b) Electron microscopy picture showing a portion of a typical human chromosome after depletion of its proteins [8]. Though its total length is a few centimetres, thermal fluctuations tend to entangle the molecule so that the typical length of a DNA coil is a few tens of microns.\]
Fig. 1.3 Schematics of the various stages of DNA compaction in the chromosomes of eukaryotes. The DNA is wrapped twice around a positively charged protein, known as a histone, to form the basic “bead on a string” structure of chromatine. This structure is further coiled into the so-called 30 nm chromatine fibre, which undergoes various poorly understood levels of supercoiling and looping to form the compact metaphase chromosome structure shown at the bottom. Adapted from [9].

1.3.3 RNA structures

Whereas double-stranded DNA is a rather simple linear polymer chain, which has a dynamical tertiary structure and is well described by models of polymer elasticity discussed in later sections, RNA (and to some extent single-stranded DNA) is capable of forming complex stable 3D structures. These structures are stabilized by base-pairing interactions along the RNA chain and are essential for the function of some RNA molecules (such as transfer RNA (tRNA) and ribosomal RNA (rRNA), ribozymes, riboswitches, etc.). Indeed, it is by now clear that the cellular role of RNA is not just to serve as a disposable copy of information encoded in DNA (the principal role of messenger RNA (mRNA), but also to act as a major regulatory and catalytic unit in the cell—a role long thought to be played only by proteins encoded in the mRNA. In fact, the 5′-untranslated region (5′-UTR) at the end of mRNA often plays a regulatory role in
transcription of DNA into mRNA or translation of mRNA into protein catalysed by the ribosome. The structure adopted by this portion of RNA (known as a riboswitch) can be altered by the binding of small metabolites, thus regulating (up or down) the mRNA translation into a protein involved in the metabolic pathway.

Ribosomal RNA (rRNA) is also an essential player in the ribosome [10, 11], not only by ensuring proper positioning of the tRNAs but also by catalysing elongation of the peptidic chain. The role of proteins in the ribosome complex is mostly structural—to serve as a scaffold for the catalytic unit formed by the rRNA.

Finally, small RNA molecules (siRNA, miRNA, etc.) play a major regulatory role in gene expression by mechanisms which are still not fully understood. This dual role of RNA, both as a support of genetic information and a catalytic molecule, has led to the suggestion that life may have originated in a purely RNA world.

From a theoretical point of view, the determination of RNA secondary and tertiary structures is somewhat simpler than the equivalent problem for proteins. First, there are fewer nucleotide types (essentially four) than amino acids (twenty). Second, their interaction is much better understood. Besides the usual self-avoidance and electrostatic repulsion, they can form Watson–Crick base pairs (and other known types of pairings, such as Hoogsteen base pairs) which can be stabilized by stacking interactions. Due to these interactions, the secondary structure of RNAs consists in a nested configuration of hairpin loops, see Fig. 1.4. This simple picture is complicated by the possibility of interactions between nucleotides on different hairpins (so called pseudo-knots).

Fig. 1.4 Secondary structure of a small RNA (the 5S-rRNA associated with the 5S subunit of the ribosome). Notice the configuration of nested hairpins characteristic of RNA secondary structures: the paired sections, labelled (2,3) and (4,5), are part of secondary hairpins in the overall 5S-rRNA hairpin which is composed of paired sections (1–5).
Nonetheless, the prediction and design of RNA structures is better understood [12] than in their protein equivalents.

As we shall see later, by allowing one to probe the structure of single RNA molecules, single-molecule techniques are ideally positioned to unravel the mechanisms of RNA folding and structural switching.
2

Manipulating DNA

The first step in any DNA manipulation experiment is to anchor appropriately labelled DNA to pre-treated surfaces (glass slides, beads, micropipettes, etc.). Several methods have been developed to achieve specific DNA binding to surfaces. They have found useful applications, from gene mapping, sequencing, and analysis [13] to the development of very sensitive immunological assays (immuno-PCR) [14]. Most of these applications achieve the required binding specificity via biochemical reactions between a labelled DNA molecule, as described in section 2.1, and an appropriately treated surface. For example, the extremity of the molecule can be functionalized with biotin (or digoxigenin) which can interact specifically with streptavidin bound to a surface (or an antibody to digoxigenin) [15]. Similarly, surfaces coated with oligonucleotides can be used to hybridize to the complementary extremity of DNA molecules. Finally, there exists a large range of chemical methods to anchor (with various degrees of specificity) the extremities of DNA to surfaces bearing reactive groups (e.g., primary or secondary amine, carboxyl, or thiol moieties).

2.1 Tailoring and labelling DNA

Compared to the chemical synthesis of artificial polymers (whose length and sequence are poorly controlled), the biochemical synthesis of DNA is a marvel of accuracy and control. Due to the importance of DNA as the support of genetic information, nature has evolved a whole panoply of enzymes which act on DNA and perform a number of “editing” functions: copy (DNA polymerases), cut (restriction enzymes), paste (DNA ligases), and correct spelling (mismatch-repair system). Although the “letters” typically used are A, G, C, and T, other “letters” are also found in nature: e.g., uracil (U) and thymine (T) in messenger RNA, or methylated cytosine (C) and adenine (A) in DNA. Artificial “letters” have also been introduced by researchers for specific purposes: e.g., labelling the four bases with different fluorescent groups for sequencing [16] or with certain chemical moieties (biotin, digoxigenin (DIG)) for specific binding to certain biomolecules (streptavidin or an antibody against digoxigenin (anti-DIG)).

The availability of these powerful molecular editing tools gives scientists unprecedented control over the synthesis of DNA, which allows them to tailor DNA sequences