Chapter 1

DNA: a dynamical object

This first chapter will be devoted to a general introduction to DNA. In Section 1.1 we will describe in a simple way the main features of DNA structure, function and dynamics. In Section 1.2 we will focus more precisely on the transcription process and we will summarize some recent biological observations on its activation and functioning. We do not attempt to be exhaustive: the aim of this introduction is just to give the necessary biological elements to understand the work presented in this text. It is very important in fact that the reader have in mind some central features of DNA structure and dynamics, as well as their biological relevance, even if it is from a quite general point of view. For this reason, we will privilege, in some cases, the simplicity of description, at the expenses of its accuracy.

1.1 An introduction to DNA

1.1.1 Structure of the DNA molecule: a helicoidal ladder

The desoxiribonucleic acid, or *DNA* [18], is a very long polymeric macromolecule, made by two coiled strands that form the well known double helix. Each strand is a chain of *nucleotides*[1]. A nucleotide is formed by a sugar (desoxiribose), a phosphate and a base. *Sugar-phosphate groups* are identical in each nucleotide and form the external *backbone* structure of the strand, being linked in succession by *covalent bonds*. A *base* is connected to each sugar to complete the nucleotide. It can be of four different types: the two purines, adenosine (A) and guanine (G), that are double-ring molecules, and the single-ring pyrimidines, cytosine (C) and thymine (T). Their chemical structure allows the bases to bind by *hydrogen bonds* in specific pairs: A with T, by a double hydrogen bond, and C with G by a triple one. Base sequences on the two strands are organized so that each base on one

[1]Some hundred in viruses, up to several thousand millions in higher organisms, with an overall length of meters (human) or even of about a kilometer (salamander). Note that such a giant molecule has to be stored in a micrometer size space, in each cell.

strand is linked to the corresponding base on the other (See Figure 1.1). The so formed base pairs are located in the core of the molecule and their *hydrogen bonds* keep together the two sugar-phosphate strands in the double helical structure. Base pairs are quite flat, and can be visualized as the steps of the DNA ladder.

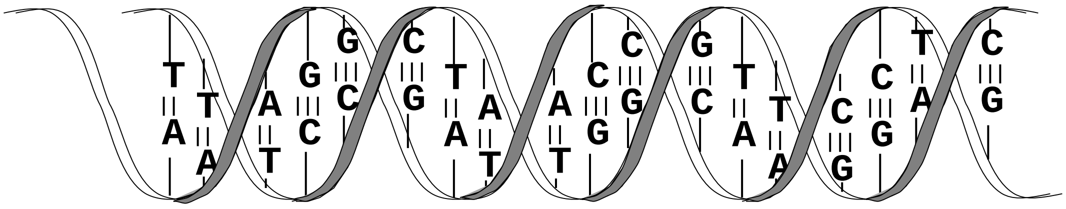

Figure 1.1: Sketch of the DNA structure. The external ribbons represent the sugar-phosphate backbone. Base pairs represent the transverse "steps" of the helicoidal ladder.

The bonds between sugar-phosphate groups on each strand allow a certain mobility. Therefore, the strands are quite flexible, but this flexibility is strongly reduced by the three-dimensional helicoidal structure that constrains the relative positions of the two strands. The double helix can adopt different conformations, among which the most common are the A and B-DNA chains. The two helices are right-handed. They differ essentially in the positioning of the bases with respect to the central molecular axis and in the inclination of the base pair planes with respect to the same axis. In the B form, to which we will refer in this work, each base pair is rotated of about 36° around the molecular axis with respect to the previous one, with the consequence that there are about ten base pairs per helix turn. The distance between neighboring base pairs along the molecular axis is about $3.4\,\text{Å}$ (Figure 1.2) [19].

There are some other right-handed forms of the helix that DNA could take depending on external conditions such as relative humidity, salt concentration, external constraints, or its specific base sequence. Left-handed helices also exist. For example Z-DNA, which is less common than A- and B-DNA, is left-handed and has a particular, zig-zag configuration of the strands. Non helical conformations are also possible such as cruciforms, hairpins or triple helix.

One of the main interests of the alternative molecular configurations mentioned above arises from the fact that they often characterize special regions in the three-dimensional spatial organization: in the living cell, DNA is in fact coiled into special hierarchical three-dimensional conformations, depending on the various phases of cellular life. It has normally fixed ends, either attached on a proteic matrix, or closed one onto the other in the case of circular DNA [20]. Non helical DNA and Z-DNA often characterize regions with special spatial locations, as *e.g.* regions where DNA is attached to the proteic matrix. For this reason functional roles have been hypothesized for these non-standard conformations.

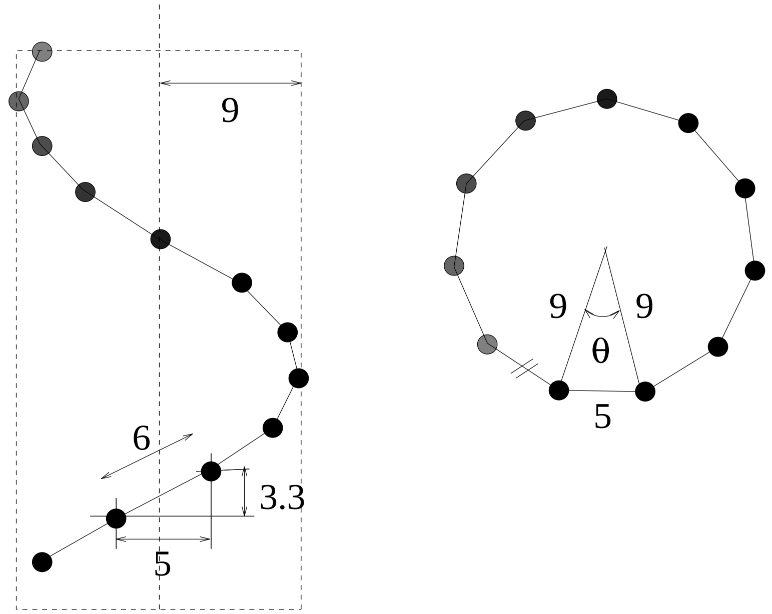

Figure 1.2: Main geometrical quantities in the B-DNA structure (lengths in Å).

Then the main interactions between nucleotides are the covalent bonds between nucleotides along each strand and the weaker hydrogen bonds between bases of corresponding nucleotides on the two strands. These are the bonds that form the ladder, which represents the *primary structure* of the molecule.

The helical shape of the molecule is referred to as its *secondary structure*. It depends on the competition among the various interactions acting between nucleotides and on the interactions of nucleotide components with the external environment, that is, in vivo, an aqueous solution: it is important to understand *why* the DNA ladder coils in the helical conformation.

Besides bonds between nucleotides, one has to consider the *base stacking* interaction and the *backbone rigidity*.

The first interaction arises from the fact that bases are hydrophobic, while sugars and phosphates are hydrosoluble. Then bases tend to stay away from water: they are tuck on the core of the molecule, while the backbone tends to stay outside "protecting" them with the external sugar-phosphate parts [19]. For the same reason, the molecular conformation at equilibrium tends to eliminate water from the core of the molecule by bringing the neighboring base-pair planes closer. Now, bases along the same strand are connected by a sugar phosphate backbone segment.

Because they are quite rigid, they separate the external ends of the base pairs by essentially a fixed length. To bring base pairs closer, it is necessary to incline them: this is obtained by inclining the strand segments in the double helix conformation.

Moreover, the molecular configuration of base pairs is caracterized by external electronic clouds, negatively charged. This gives rise to an electrostatic repulsion between neighboring base pairs, which stabilizes the helical conformation: it prevents base pairs from getting closer and "selects" the rotation angle that corresponds to a minimum repulsion.

Furthermore, it is interesting to mention that there are several other electrostatic interactions between specific molecular sites and the solution components that contribute to stabilize the secondary structure. For example, sites on opposite strands which are half pitch of the helix away in the sequence, and which result therefore nearby in three-dimensional space, are connected by water bridges forming filaments [21].

1.1.2 Functions of the DNA molecule: protein synthesis

The cellular structures are mainly made of *proteins*, and proteins mediate most of the biological processes involved in cellular life. The two complementary sequences of bases lying on the strands have the role of coding for protein synthesis. Proteins are chains of *amino acids* (typically several hundreds), and the specific sequence of amino acids, which are of 20 different types, determines the protein shape and function: DNA provides the instructions for building all needed proteins by connecting amino acids in the right order. This is acheived by the *genetic code*, that associates to each different amino acid a group of three bases called *triplet*, or *codon*. A DNA *coding sequence* is formed by a sequence of such codons, that represents exactly the amino acid chain of a given protein. We denote *gene* the DNA segment that contains the complete coding sequence for a given protein[2]. It can be read by a special cellular machinery that is able to construct the protein following the encoded instruction. By simplifying extremely, we can summarize the complex process leading to protein synthesis by its two main steps:

1. The sequences of codons contained in DNA are first read and copied in a complementary *RNA* (ribonucleic acid) chain by an enzyme called RNA-polymerase. This phase is called *transcription process*.

2. The obtained RNA chain is in turn read by *ribosomes* that are able to recognize and "catch" amino acids dispersed in the cellular cytoplasm and to join them following the order encoded in the RNA instructions. This is the *translation process*.

DNA contains also the instructions that allow the selective activation of the different proteins synthesis. In order to produce a specific protein when and where it is needed, a whole series of inhibition/activation mechanisms is engaged

[2]A gene can anyway contain also additional non coding parts, that will be accurately cut away before translation (RNA splicing or maturation).

throught enzyme and protein interactions with some special DNA sites called promoters and enhancers (We will consider this point more precisely in Section 1.2). These regulatory regions do not contain protein coding regions. Their base composition (often characterized by repeated sequences of different types) is probably related to structural properties such as flexibility, intrinsic bending or hydrogen bond strength, which influence the steric interaction with binding proteins.

1.1.3 Dynamics of the DNA molecule: transcription and denaturation processes

In the present work, we will focus in particular on the dynamical characterization of the transcription process. As we already know, the main function of DNA is to contain the coded instructions for the protein synthesis. We also know that the bases are the fundamental coding units, and that they are coupled in pairs; furthermore, in the equilibrium DNA configuration base pairs are enclosed in the molecule core. To perform transcription, the coding sequence on one strand has to be exposed to the exterior of the molecule to be read by the RNA-polymerase and copied into RNA. It is clear that, to allow base sequence reading, hydrogen bonds must be temporarly broken and the two strands have to separate. As all the other DNA biological processes, transcription involves obviously some *dynamical modification* of the molecular structure.

Overall, each nucleotide is composed of about thirty atoms, that are individually characterized by small amplitude, fast vibrational motions (on scales of $0.01 \div 0.1$ Å and $0.01 \div 0.1\,ps$) superimposed on larger and slower motions of atom groups [22].

Between the several kinds of motion that could take place in a DNA segment, the separation of bases in a pair (with hydrogen bond breaking) has a particular interest, being involved in transcription functioning, *i.e.* in the DNA most important activity[3]. During transcription, a region of about 15-20 base pairs (*bps*) opens and forms the so called *transcription bubble.* Base pair opening is also called *denaturation.*

Base pair opening could also be induced by heating DNA in solution (*thermal denaturation* or *melting*). At lower temperatures, the thermal energy induces oscillatory motions of base pairs on the one site potential wells, with a relatively low frequency and long life time, with an overall motion that has been called *"breathing"*.

Breathing modes have been revealed by NMR studies, showing in the DNA the presence of large conformational fluctuations with time constants of the order

[3]As well as in DNA replication, when the two strands have to separate completely to give rise to two new complete DNA double chains to be transferred to the daughter cells.

of about $0.3 \div 0.7$ *psec* [23].

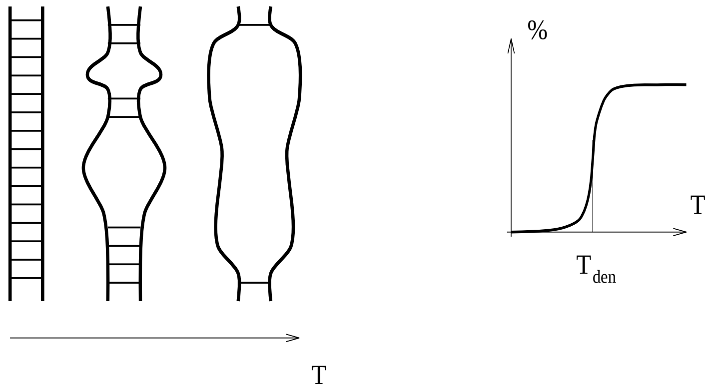

Figure 1.3: Schematic representation of thermal denaturation process. **(a)**: Bubble formation; **(b)**: denaturation curve: this curve shows the typical evolution of the percentage of denaturated base pairs as a function of the temperature. T_{den} is the transition temperature.

These localized oscillations correspond to local alternate hydrogen bond breaking and reclosing, *i.e.* to small denaturated bubbles that oscillates. When the temperature increases, the bubbles grow and then combine together to form bigger bubbles until the complete separation of the two strands. This arises around a typical transition temperature T_{den}. The double helix stability, *i.e.* its resistance against denaturation, depends on its sequence: an AT-rich sequence, *i.e.* a sequnce which contains a large percentage of AT base pairs, is less stable, because the AT pairs are linked by double hydrogen bonds, easier to break than CG pairs triple ones. The denaturation temperature T_{den} is in fact a good indicator of the AT/CG content ratio of a sequence. Denaturation curves are obtained by measuring the UV absorbtion of a DNA diluted solution as a function of slowly increasing temperature (*cf.* Figure 1.3). The heterocyclic rings of nucleotides absorb in fact light in the ultraviolet range, with a maximum close to $260\,nm$. But the absorbtion of DNA itself is some 40% less then for free nucleotides (in solution), beacause of an effect depending on interactions between the electrons of the bases which relates to the stacking interaction in the double helix conformation [18]. The optical density indicates then the open pairs percentage.

The melting temperature T_{den} is normally in the range of $326 - 370K$, and denaturation occurs in a very narrow temperature range of a few degrees.

1.2 Transcription initiation processes and double helix structure.

We will go now into a more detailed description of the transcription process. This is the process by which the instructions coded in a gene can be read and transcribed on a RNA chain, then used as a template for the synthesis of proteins. Many different mechanisms cooperate in the different phases of this complex function, regulating their efficiency rates. The whole process remains, at present, not well understood in many aspects.

Nowadays, experimental biology gives clear evidences of the presence of a strong correlation between *structure* and *function* in most DNA chemical processes. A great part of the current biological research is devoted to the study of structural mechanisms involved in DNA biological functions, in order to understand their direct effects on the processes involved. In this section, we will discuss some complex interactions occurring, during transcription, among the different structural modifications involved, and the effects of these interactions on the transcription efficiency itself. The pourpose of this section is to show that modifications of the torsional properties of the helix, namely twist angle modifications, are crucial for transcription initiation.

1.2.1 Transcription initiation process in details. DNA opening

Transcription (Figure 1.4) is a complex process that involves a large number of enzymes. The main enzyme responsible for transcription, the *RNA-polymerase*, has first to recognize a specific region in the DNA chain, called *promoter*, that is located immediately upstream with respect to the gene, and to bind to it, forming the *initiation complex* [19, 20]. In most cases there are some other proteins or *transcription factors* that bind to the promoter together with the RNA-polymerase: they are needed to maximize the transcription rate, or, in some cases, to allow the transcription itself.

The next step is the formation of the *open complex*. A small segment (of about 15-20 *bps*) of the helix in correspondence to the promoter region is "melted": the two strands are separated forming the *transcription bubble*. It is interesting to note that, in this phase, no chemical energy is required by the RNA-polymerase to open the two strands. The mechanism that allows the breaking of the hydrogen bonds in the bubble is still a puzzling problem.

Then the RNA-polymerase and the transcription bubble start to move along the gene, copying the coding sequence into RNA (*elongation phase*).

The formation of the open complex and its motion along the chain are often regulated by a set of activation/inhibition processes. They function by means of the binding of other proteins to different DNA regions that could be very far

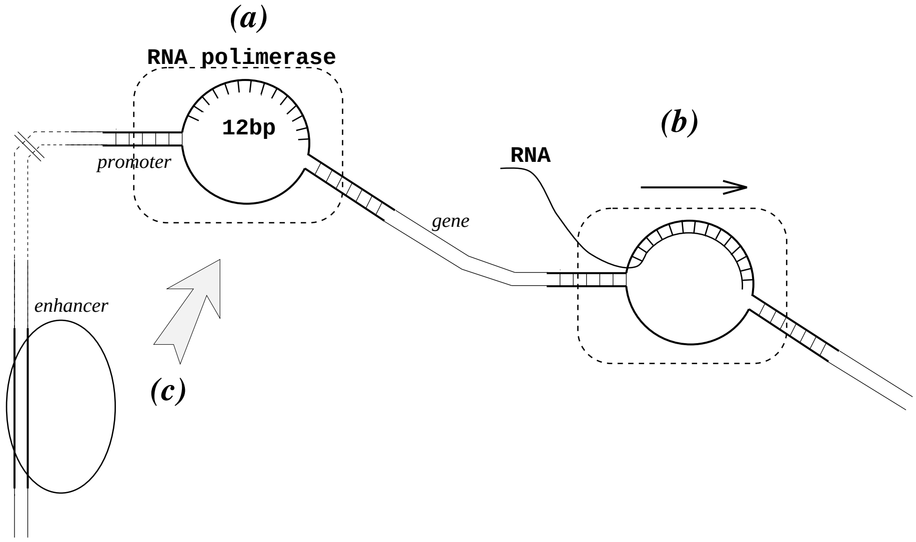

Figure 1.4: Schematical representation of the transcription process. **(a)**: open complex formation; **(b)**: elongation phase; **(c)**: enhancer linking protein activation.

along the chain either upstream or downstream with respect to the promoter. This starting activating sites can be upstream and close to the promoter region (usually with a distance along the chain of about 100-200 *bps*); they are called *enhancers*, when they are instead far from it (up to several kilobases). These mechanisms of activation/inhibition are still mostly unknown, particularly for what concerns the latter.

We shall consider in particular transcription initiation, *i.e.* the group of processes that allow the transcription to start.

1.2.2 Observations and hypothesis for the dynamics of transcription activation

To explain the long range effect in transcription activation many alternative models have been proposed. Electron microscopy has allowed the visualization of one of the mechanism of activation for transcription enzymes bound to DNA sites distant from the promoter [24]. It has been observed that they loop the molecule so to bring the promoter and enhancer regions closer (Figure 1.5(a)) [25]. According to this result, it is natural to think that the principal mechanism is a direct interaction between the various bound proteins.

However the formation of a loop is not the unique way for activation, and two other mechanisms for these long range activation effects has been proposed [26]: first, the induction of the cooperative binding of transcription factors, so that all the DNA between enhancer and promoter is completely covered of bound proteins (*"oozing"* model, Figure 1.5(b)). Second, the transmission of altered

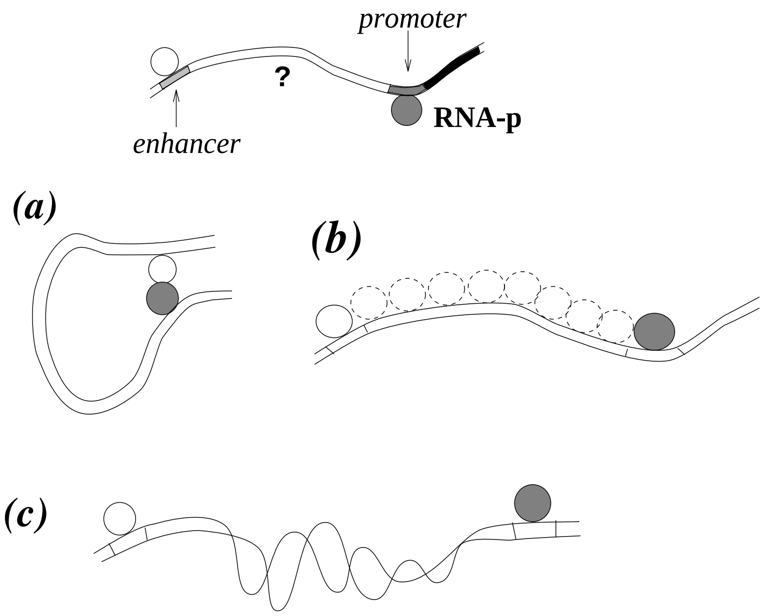

Figure 1.5: Proposed enhancer activation mechanisms: **(a)**: loop formation; **(b)**: "oozing"; **(c)** distortion transmission.

DNA structures from the enhancer, along the helix, to the transcription complex (Figure 1.5(c)) [20].

Among these two proposed mechanisms, the latter seems to correspond better to experimental results. It is in fact known that the molecular conformation of the enhancer is modified by the linking proteins: for example, the enhancer region is bent by these activator factors. Furthermore, experimental works seem to confirm that the structural modifications induced in the enhancer sites are actually much more important, for activation, than the specific (chemical) properties of the linked proteins that mediate activation. In some cases in fact proteins with a strongly different chemical structure are able to activate the same process, provided that the same modifications are induced into the DNA structure. It has also been shown that activation can be achieved, in some cases, by replacing the enhancer region by an intrinsically bent [27] or superhelical [28] DNA sequence that is not a protein binding site, so that there is no more protein mediation: in this case, thus, it is necessarily the structural deformation which acts as activator. We stress anyway that the various models of action at a distance are often interrelated, *e.g.* because the formation of loops necessarily involves topological and structural changes that modify the molecular helicity and bending properties [26].

It is interesting to mention also the proposed *"hit and run"* mechanism for DNA binding proteins, according to which, immediately before the open complex formation, activator factors bind just for a short time to DNA, locally modify its structure, and then leave the binding site [29]. This could suggest that the action of these activator factors could be that of "launching" a structural distortion that could travel along the chain towards the promoter region, where it will eventually act as activator by inducing the right conformational change.

1.2.3 Structural deformation as the central feature for DNA opening

These are not the only cases in which structural deformations can be shown to be relevant for biological functioning in transcription (in long range activation effects as well as in open complex formation) and bending is not the unique kind of "active" deformation. The winding of the two strands of DNA around one another in the double helix has some crucial consequences for its functions. Besides the cited results there are for instance several evidences of the enhancing effects of intrinsically superhelical sequences, *i.e.* regions where the rotation between the neighboring base pair, or *twist* angle (Figure 1.6), differs with respect to its value in equilibrium conditions [30, 4]. If the double helix is twisted in the opposite sense with respect to the windings of the two strands, the resulting torsional force can be relieved by partially unwinding the two strands. If the torsion is great enough, it may even lead to a limited disruption of base pairing [18]. The twist deformations are in fact strongly related, for geometrical reasons, with the bubble formation. This kind of effect has, from our point of view, a great interest.

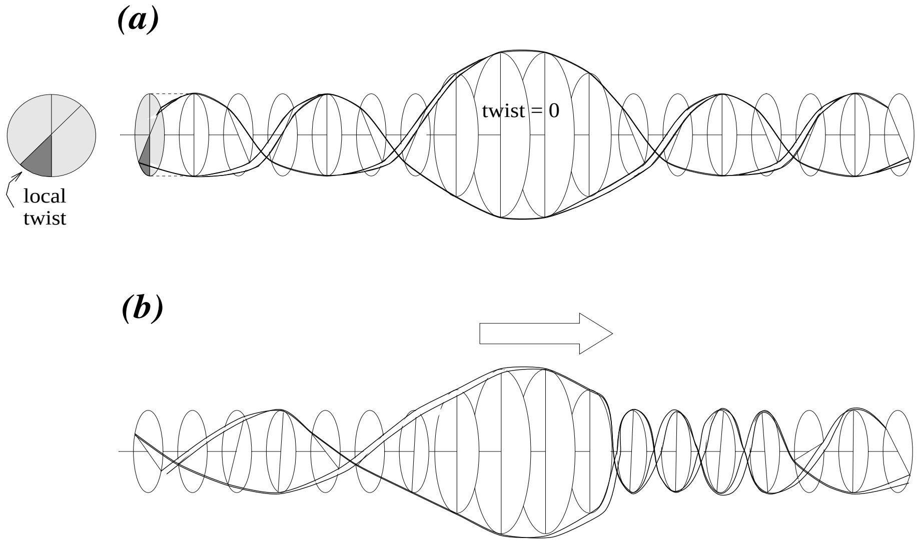

Figure 1.6: Local twist effects. **(a)**: the bubble local opening is possible only if accompanied by an untwisting; **(b)**: in the elongation phase, the motion of the untwisted bubble along the helix produces some overtwist upstream and some undertwist downstream.

Moreover, local changes in twist are necessarily and directly involved in the transcription bubble formation: simple geometrical considerations allow to understand that the stretch of the hydrogen bonds in some local region of the chain is possible only if the twist angle is decreased with respect to its equilibrium value Θ_0 in that region, in order to bring the steps of the double helix "ladder" toward a common vertical plane (Figure 1.6(a)) [19]. This constraint is well known by biologists because in the elongation phase, when the bubble starts to move, it causes topological problems, giving rise to a positive excess of twist in front of the RNA-polymerase and to a negative excess behind (Figure 1.6(b)). Special enzymes, the *topoisomerases*, are designed to release these structural stresses by cutting one of the two strands and rejoining it after having made some turns around the other. If the topoisomerases cannot act efficiently enough, the excess of twist could sometimes be sufficiently important to prevent transcription, showing how the twist deformations are important in all DNA opening processes.

1.2.4 Supercoiling effects

In DNA, changes in the twist angle can always be put in relation with modifications in its three-dimensional shape.

The DNA *tertiary structure* is the way in which the long double helix is organized in the three-dimensional space. The definition includes the complex hierarchical organization of DNA in living cells, but also the simpler structures formed in space by a DNA segment under certain geometrical constraints. In particular, if a torsional constraint is imposed, *e.g.* in the case of circular molecules, the main molecular axis tends to form a solenoid or to coil onto itself in a interwound structure called *plectoneme* (See Figure 1.7).

This phenomenology is usually referred to as *supercoiling* [18, 19]. The principle behind supercoiling is that the helix cannot unravel when it is wound around itself and its ends are fixed. It is possible to relate the DNA local angular variables (twist) with its three-dimensional structure by the law:

$$Lk = Tw + Wr\,, \tag{1.1}$$

where Tw (twisting number) is the number of turns that one strand makes around the other, and is then the ratio between the overall twist angle change along the considered DNA segment and 2π; Wr (writhing number) is the number of times the molecular main axis crosses over itself, giving rise to a coiled three-dimensional configuration; Lk (linking number) is the total number of times the two strands wrap around, and is a fixed number in closed circular chains [19]. Many recent works are devoted to the study of supercoiled DNA structure[4] their influence on transcription and on other processes.

[4] Just as examples we can cite, among many others, Refs. [1]-[6]

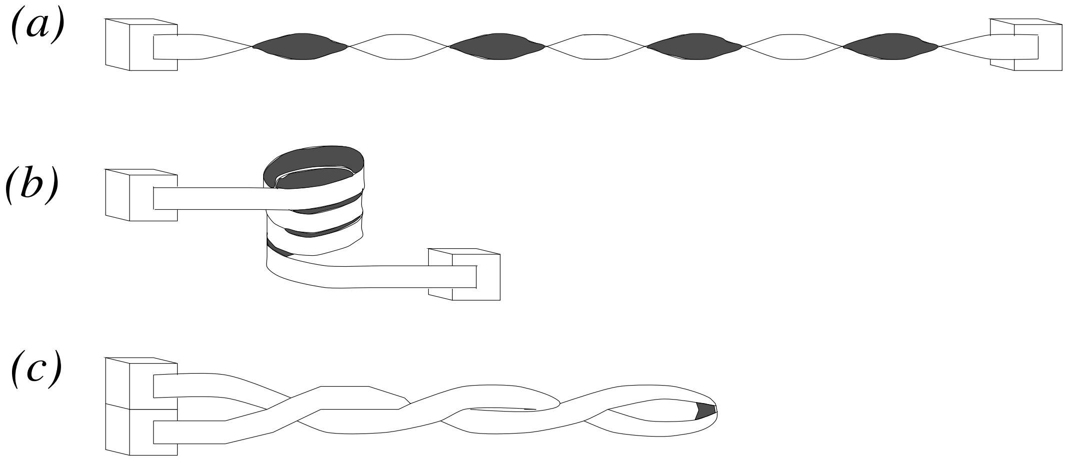

Figure 1.7: Supercoiled Structures. **(a)**: a twisted, straight structure, in which the excess linking number is all contained in the twist distortion, and $Wr = 0$; **(b)**: a supercoiled structure of the solenoidal type in which $Tw = 0$ and $Lk = Wr$; **(c)**: a supercoiled structure with again $Tw = 0$ and $Lk = Wr$, but forming a *plectoneme*. We stress that, in these pictures, the DNA chain is represented as a flat ribbon, without reproducing its intrinsic helical structure.

1.3 Studying the transcription initiation by modeling DNA structure

Transcription initiation is a complex process, not well understood. From what we have discussed up to now, it turns out that it is in many aspects strongly dependent on structural modifications.

Based on a physical point of view, our aim is to to build a helpful theoretical tool to study some physical features implied in the actual mechanisms of transcription initiation (and related features such as transcription bubble formation and denaturation).

This can give elements for answering the following questions:

1. how can RNA-polymerase collect enough energy in the promoter region to melt the DNA in a bubble?

2. Can we better describe the structural mechanisms which are responsible for the activation induced by the enhancer linked proteins?

3. Is it possible to have traveling distortions in a helical structure? Which should be their shape?

4. Is it possible to model the mechanism through which base pairs opening is influenced by the specific helical properties of the chain?

To do this, it is necessary to simplify our description of the double helix, in such a way that the problem can be treated in a mathematical framework. We will take particular care to include in this scheme the most essential geometrical constraints of the molecule to have a quite realistic description of the main dynamical features of the various processes involved. This can be achieved by DNA *dynamical modeling*, *i.e.* by describing its structure and motion through a simplified set of its most relevant properties.

The kind of motion involved in denaturation and transcription implies the displacement of the entire nucleotide: it can be treated through a mesoscopic simplification of the molecular description which reduces the nucleotide groups into single objects. An appropriate set of degrees of freedom has to be chosen. This can be achieved by taking into account the indications arising from biologically observed properties described in the present chapter. In Chapter 2, will describe how to choose a simple set of degrees of freedom and we will introduce the essential interactions in order to build a DNA model which allows such a schematic description of the molecule.