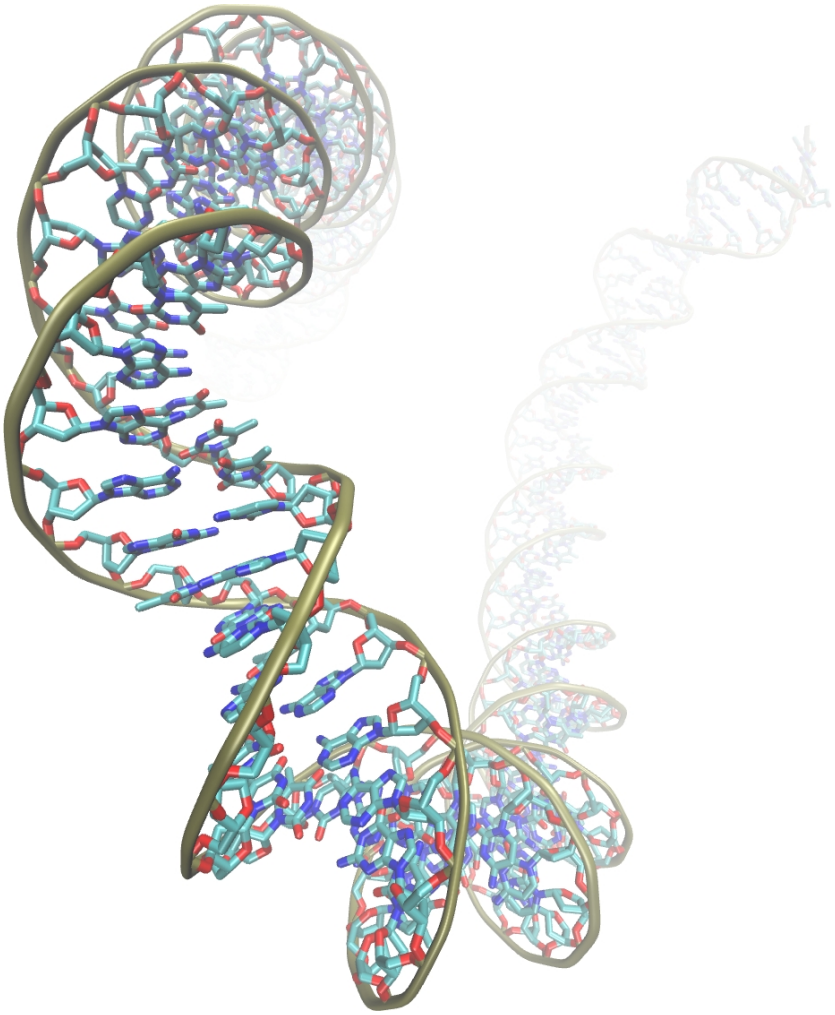
Case Study: DNA

Leonardo Trabuco and Elizabeth Villa



1 The Molecule of Life

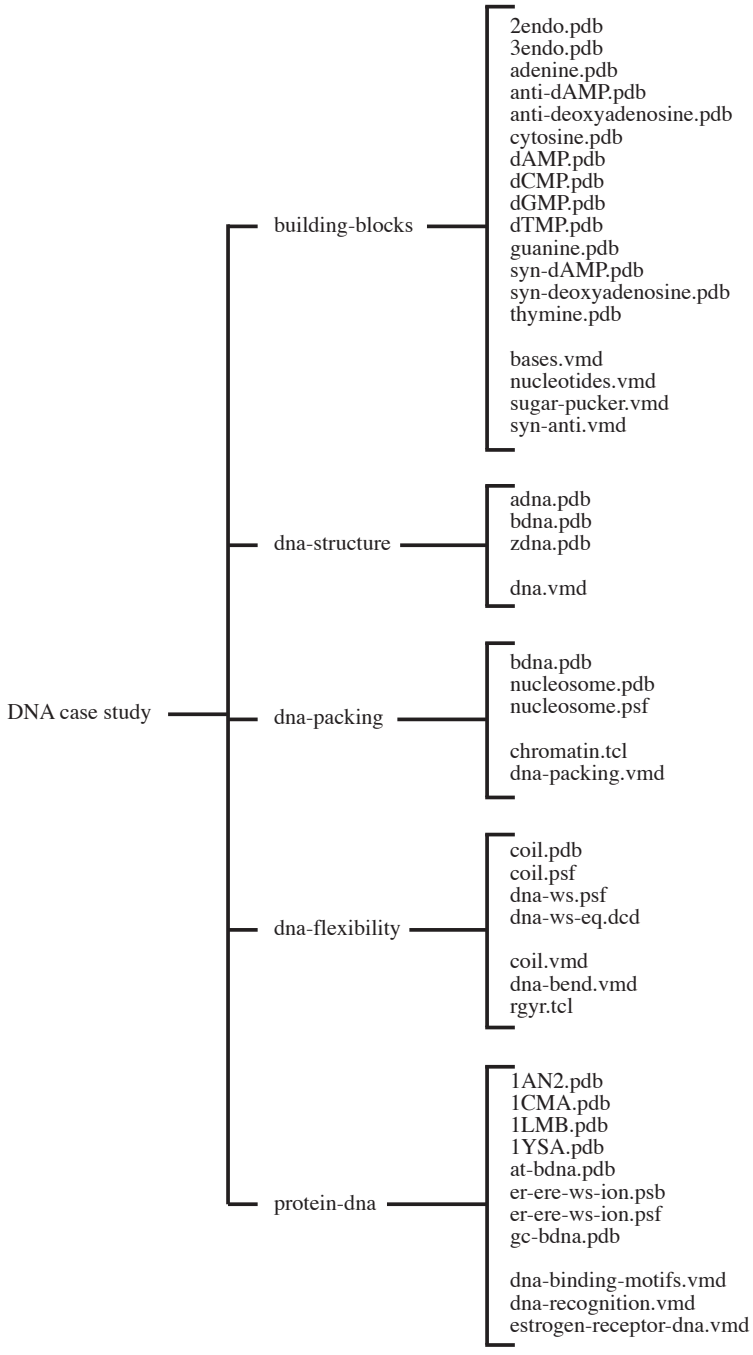
With the 50^{th} aniversary of the discovery of the DNA structure by James Watson and Francis Crick and the race for the human genome, with all its controversies and awe-inspiring medical implications, we have all heard a lot about DNA, the molecule of life.

What makes DNA "the molecule of life"? DNA, or deoxyribonucleic acid, is the molecule that carries all the genetic information of an organism (except for some RNA viruses). It can be thought of as a blueprint containing the instructions that govern the production of proteins and other molecules essential to cell function. The collections of these instructions is called a **genome**. The informational units of the genome are called genes. Genes are translated into protein via the genetic code, which defines the protein sequence. This translation process is central to life.

Despite being so important, DNA in reality is a passive molecule. As an instruction book, all DNA does is to store information. DNA is therefore like any other data-storing device: it needs to be read, stored, copied and otherwise manipulated by other biomolecules. These other macromolecules exist in different forms and types, and interact with DNA in different ways, e.g., by bending it, copying it, denaturing it, packing it, walking on it, nicking it, repairing it, etc. All this is done in order to preserve and obey the set of instructions that give life to the organism. In this sense, there is no life without genetic information.

As passive as it might be, the DNA molecule bears incredible physical properties that allow for long genomes to be stored in compact nucleii and for extreme manipulation resulting in replication and gene regulation. In this case study, we will introduce you to this wonderful molecule of life. In section 2, you will learn about the chemical components that form DNA. Section 3 will introduce you to the different helical shapes that DNA adopts. In section 4, you will learn of the tight packing that DNA needs to undergo in order to be stored in the nucleus; the flexibility that DNA must have to make this possible will be explored in section 5. Finally, section 6 will introduce you to some of the proteins that manipulate DNA and illustrate the mechanisms behind protein-DNA interaction.

The figure below lists all the files that were provided with this case study; you will need them to complete the exercises and to explore further the molecules presented in the figures.


DNA case study
building-blocks
2endo.pdb
3endo.pdb
adenine.pdb
anti-dAMP.pdb
anti-deoxyadenosine.pdb
cytosine.pdb
dAMP.pdb
dCMP.pdb
dGMP.pdb
dTMP.pdb
guanine.pdb
syn-dAMP.pdb
syn-deoxyadenosine.pdb
thymine.pdb
bases.vmd
nucleotides.vmd
sugar-pucker.vmd
syn-anti.vmd
dna-structure
adna.pdb
bdna.pdb
zdna.pdb
dna.vmd
dna-packing
bdna.pdb
nucleosome.pdb
nucleosome.psf
chromatin.tcl
dna-packing.vmd
dna-flexibility
coil.pdb
coil.psf
dna-ws.psf
dna-ws-eq.dcd
coil.vmd
dna-bend.vmd
rgyr.tcl
protein-dna
1AN2.pdb
1CMA.pdb
1LMB.pdb
1YSA.pdb
at-bdna.pdb
er-ere-ws-ion.psb
er-ere-ws-ion.psf
gc-bdna.pdb
dna-binding-motifs.vmd
dna-recognition.vmd
estrogen-receptor-dna.vmd

2 The Building Blocks

The DNA molecule is a polymer composed of nucleotides. Each of these nucleotides is a letter in the "genetic alphabet". Each nucleotide is formed of a phosphate group, a sugar and a nitrogenous base (Fig. 1). Two nucleotides are connected by a phosphodiester bond, in which a phosphate group links the 3'-hydroxyl group of one nucleotide to the 5'-phosphate group of the next, giving rise to directionality in the polynucleotide chain. By convention, we refer to DNA sequences from the 5' to the 3' end, which corresponds to the direction in which they are synthesized and read by other molecules.

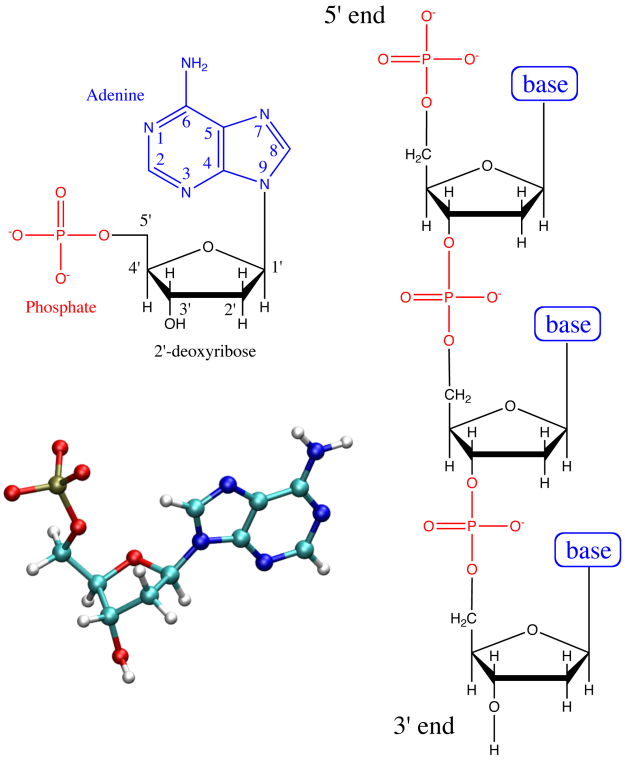


Figure 1: Each nucleotide in DNA is composed of a phosphate group, a sugar (2'-deoxyribose), and a nitrogeneous base (adenine, in this case). On the right you can see how the directionality of the polynucleotide chain comes about. You can explore the structures of all nucleotides yourself with the VMD saved state `building-blocks/nucleotides.vmd`.

The base defines the type of nucleotide. There are two types of bases in DNA: purines and pyrimidines. The two main purines that occur in DNA are adenine (A) and guanine (G). As for pyrimidines, cytosine (C) and thymine (T) are typically found (Fig. 2).

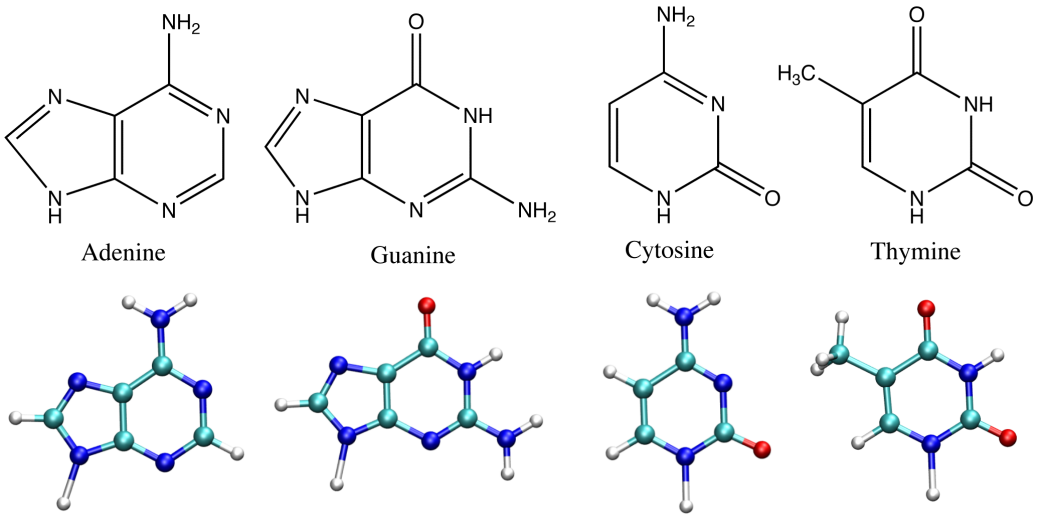


Figure 2: Nitrogeneous bases occurring in DNA: adenine (A) and guanine (G) are purines, whereas cytosine (C) and thymine (T) are pyrimidines. You can explore the structures of the bases with the VMD saved state `building-blocks/bases.vmd`.

The sugar ring is generally non-planar, and determines the structure of DNA, as we will see later. It can be puckered with one or two adjacent atoms out of the plane formed by the remaining atoms. Atoms displaced from these three- or four-atom planes and on the same side as $C_{5'}$ are called *endo*; those on the opposite side are called *exo* (Fig. 3). Besides, the base can adopt two main orientations relative to the sugar, called *syn* and *anti*, as shown in figure 4 [1].

3 Nucleic Acid Structure

As in the case of proteins, the nucleic acid structure is divided hierarchically into primary, secondary, and tertiary structures. The primary structure corresponds to the linear sequence of nucleotides, just as in proteins the primary

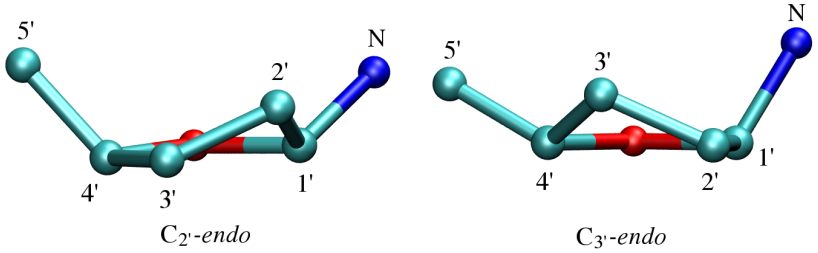


Figure 3: Main sugar puckering modes observed in DNA. You can examine the differences between these two main conformations with the VMD saved state `building-blocks/sugar-pucker.vmd`.

structure corresponds to the sequence of amino acid residues. The determination of the sequence of large nucleic acids reveals the genetic information of an organism, a task that has turned from impossible into routine in a matter of a few decades. Today, the entire genomes of hundreds of organisms from all domains of life are known, including the genomes of the bacterium *E. coli*, the fruit fly, the dog, and human. Indeed, sequencing of short segments of DNA is done routinely in many laboratories. Currently, researchers are pushing for advances in nanotechnology that will permit us to sequence entire genomes whithin days at low cost; a twenty million dollar prize has been announced for the team that pushes the cost below $1000.

The secondary structure of DNA refers to the different helical arrangements adopted by the molecule, the principal forms being described in this section. The tertiary structure of DNA corresponds to higher-order structures that it adopts in different cellular contexts, the most remarkable being chromosome packing, explained in Section 4.

3.1 The DNA Double Helix

The discovery of the DNA double helix is one of the most fascinating chapters in the history of science. DNA was first isolated and characterized by Friedrich Miescher in the 19th century [2], but it wasn't until 1944 that it was shown to be the unit of inheritance by Oswald T. Avery, Colin MacLeod, and Maclyn McCarty [3]. They found that DNA extracted from a virulent strain of the bacterium *Streptococcus pneumoniae*, genetically transformed a nonvirulent strain of this organism into a virulent form [4]. Further evidence

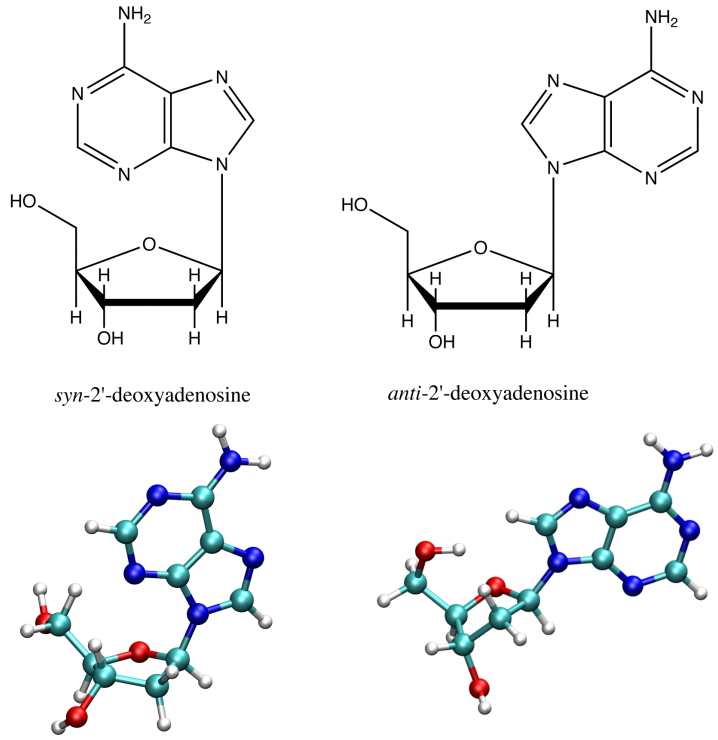


Figure 4: *Syn* and *anti* orientation about the glycosil bond for the nucleoside 2'-deoxyadenosine. A nucleoside is the same as a nucleotide without the phosphate group. You can explore the differences on the structures of the *syn* and *anti* forms with the VMD saved state `building-blocks/syn-anti.vmd`.

that the DNA carries the genetic information came from Alfred D. Hershey and Martha Chase's work, in 1952 [5]. They showed that when the bacterial virus (bacteriophage) T2 infects its host cell, *E. coli*, it is the DNA, not the protein, that enters the host cell and furnishes the genetic information for viral replication [4]. Chargaff and colleagues, in the late 1940s, discovered that adenine and thymine always appear in equal number in DNA, as do guanine and cytosine. These quantitative relationships (A=T and G≡C) became known as the "Chargaff's rules" [6].

Rosalind Franklin and Maurice Wilkins studied the X-ray diffraction pattern of DNA fibers. Without the benefit of the X-ray diffraction patterns of DNA fibers, Linus Pauling and Robert Corey proposed in 1953 that the DNA

was a triple helix [7]. Soon after, James Watson and Francis Crick published what is probably the most famous paper in biology [8], describing the DNA as a right-handed, **anti-parallel** double helix with a specific pattern of base pairing (isomorphous A=T and G≡C pairs), accounting for all experimental evidence available thus far, including Chargaff's rules (Fig. 5). Their model was purely theoretical, based on available experimental data, including access to Franklin's DNA images without her knowledge. There is evidence that she was on the right track when Watson and Crick published their model [9], which gave rise to a big controversy, especially because only Crick, Watson and Wilkins shared the 1962 Nobel Prize in Physiology and Medicine.

In their model, Watson and Crick revealed that DNA formed a double helix, with a helix rise per base pair (helical pitch) of 3.4 Å, and 10 base pairs (bp) per helical turn. In solution the structure is somewhat different from that in fibers, having 10.5 bp per turn. Two grooves with different width and depth can be identified on the surface of the helix, due to the offset between the strands, and are referred to as **major** and **minor grooves** (Fig. 5). The base-base interaction that contributes the most to the stabilization of the double helix is perpenticular to the helix axis (base stacking), stabilized mainly by London dispersion forces and hydrophobic effects [1].

The two DNA strands are complementary, and the specificity is given by Watson and Crick base pairing, which we write as A=T and G≡C in order to emphasize the fact that they form two and three hydrogen bonds, respectively (Fig. 5). The complementarity of the double helix permitted us to understand how the genetic information can be accurately replicated, which literally caused a revolution in the biological sciences.

3.2 Structural variants of DNA

After the discovery of the DNA structure by Watson and Crick, it was revealed that DNA can occur in several different structural forms. It became apparent with the first X-ray diffraction patterns of DNA fibers that two different allomorphs existed, which Rosalind Franklin termed A- and B-DNA [10]. The main difference between the A and B forms is the sugar puckering mode, $C_{3'}$-*endo* for A and $C_{2'}$-*endo* (or the equivalent $C_{3'}$-*exo*) for B form [1]. The Watson-Crick structure corresponds to what is called B-DNA, the most stable structure for a random-sequence DNA molecule under physiological conditions, whereas the A form is favoured in solutions relatively devoid of water [4] (Fig. 6).

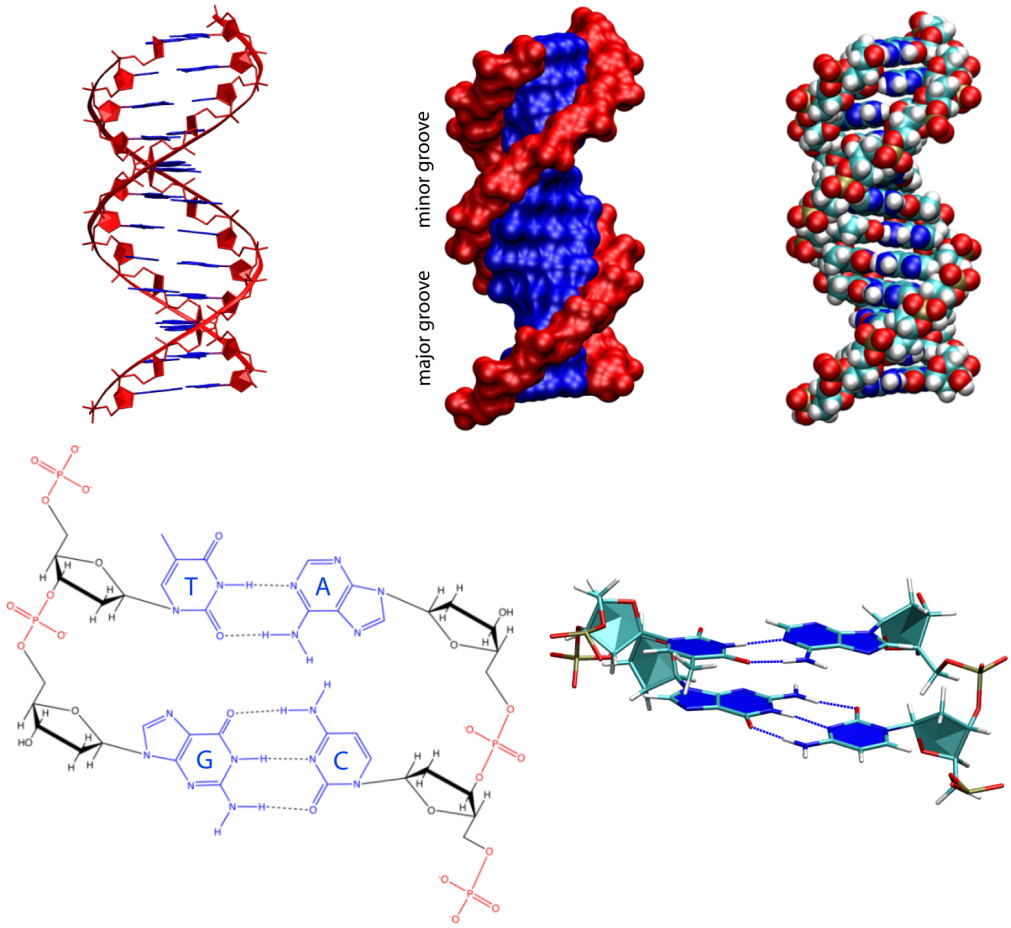


Figure 5: Three different representations of the DNA double helix model proposed by Watson and Crick (*top*). Two adjacent Watson-Crick base pairs (A=T and G≡C) are depicted, emphasizing the hydrogen bonds between the bases (*bottom*). You can explore this structure with the VMD saved state `dna-structure/dna.vmd`.

All the structural information of DNA until 1970s was derived from X-ray diffraction of DNA fibers. Interestingly, the first single-crystal X-ray diffraction revealed a totally unexpected left-handed double helical structure [11]. Due to the zigzag arrangement of the backbone it was termed Z-DNA, and only one groove is apparent, which is analogous to the minor groove in B-DNA [12]. Other characteristics are summarized in Table 1.

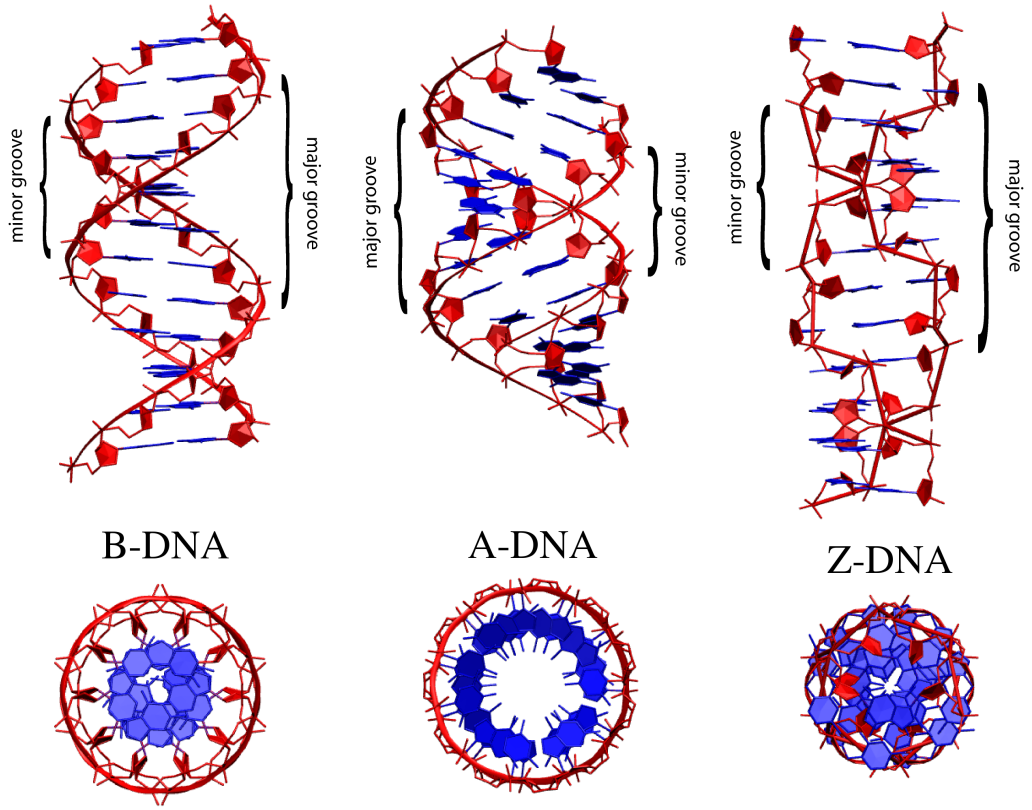


Figure 6: Structural variants of DNA. With the VMD saved state `dna-structure/dna.vmd`, you can explore the structural differences between the main DNA forms depicted above.

Z-DNA is a transient structure, stabilized by negative supercoiling which is generated both by the unwrapping of DNA from nucleosomes and by the RNA polymerase complex itself. There exist also sequences that favor Z-DNA formation which oftentimes occur in the beginning of genes. Besides, proteins that specifically bind Z-DNA have been discovered, establishing a biological role for Z-DNA [12].

	A-DNA	B-DNA	Z-DNA
Helical sense	Right handed	Right handed	Left handed
Diameter	$\sim$ 26Å	$\sim$ 20Å	$\sim$ 18Å
Base pairs per helical turn	11	10.5	12
Helical rise per base pair	2.6 Å	3.4 Å	3.7 Å
Base tilt normal to the helix axis	20 °	6 °	7 °
Sugar pucker conformation	$C_{3'}$-*endo*	$C_{2'}$-*endo*	$C_{2'}$-*endo* for pyrimidines; $C_{3'}$-*endo* for purines
Glycosyl bond conformation	*Anti*	*Anti*	*Anti* for pyrimidines; *syn* for purines

Table 1: Summary of the differences between the three main DNA double-helical conformations (from [4]).

Exercise 1: Double helix of DNA. In this exercise, you are going to explore the various structures of double helical DNA. Load the VMD saved state `dna-structure/dna.vmd`, which contains the structures of A, B, and Z-DNA[a].

1. Measure the distance between two contiguous base pairs (helical rise) in each of the structures.

2. Make a H-bonds representation (use a distance cutoff of 3.2 Å) for a G≡C and an A=T pair and create a snapshot. How many hydrogen bonds are there in each case?

3. When segments of DNA are subject to high temperature, the double helix melts. There is a difference in melting temperature between segments with high A=T and G≡C content. Which case would correspond to the higher melting temperature? Explain this difference based on Watson-Crick base pairing.

4. Create a van der Waals representation for each structure with one color for the backbone (you can use the keyword "backbone" in the selection) and another for the nucleoside ("not backbone"). Create a snapshot, identifying the major and minor grooves (see text), as well as the helical sense (right or left handed helix) for each DNA form. Measure the size of the major and minor grooves for each DNA form.

[a]The PDB files were obtained from
`http://chemistry.gsu.edu/glactone/PDB/pdb.html`

4 DNA packing inside the cell

The genetic information of a cell is stored in DNA. Each cell contains the whole genome of the organism. In prokaryotes, the DNA is simply contained in the cell; in eukaryotes, the DNA is tightly packed inside the nucleus. Typically, if one stretched the DNA contained inside a cell, its length would be thousands of times larger than the cell itself. Figure 7 shows a photograph of an *E. coli* cell placed in a solution with a high salt concentration, which caused the cell to burst and eject the DNA it contained. You can appreciate from the image that DNA needs to be packed very efficiently inside the cell.

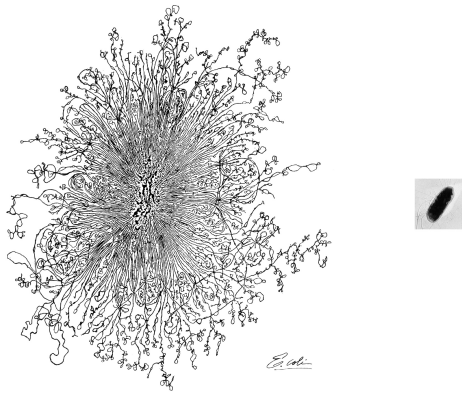


Figure 7: Photography of an *E. coli* bacterium with its genomic DNA extruded onto an electron microscope slide. *E. coli* is about 1-2 μm long. Its circular genome has 2 million base pairs, which is about 1000 times longer than the bacterium itself. The mass in the center is what's left of the bacterium after lysis [13]. On the right, an *E. coli* cell for comparison.

The DNA stored in each human cell has a total length of about 6 feet or 1.8 m; the size of a typical human cell is 10^{-5} m. The DNA inside a human cell is packed into the nucleus to occupy only 10% of the total cellular volume. For such efficient packing, DNA needs to be folded into higher-order structures. Figure 10 shows the different scales of DNA packing inside the nucleus.

The hierarchical packing of DNA is carried out by specialized proteins, the most abundant of which are called histones. In fact, there is about the same mass of histones as there is mass of DNA in a cell's nucleus. Histones are small proteins with a high positive charge (they contain many lysine and arginine residues), which helps the histones bind tightly to DNA. Eight his-

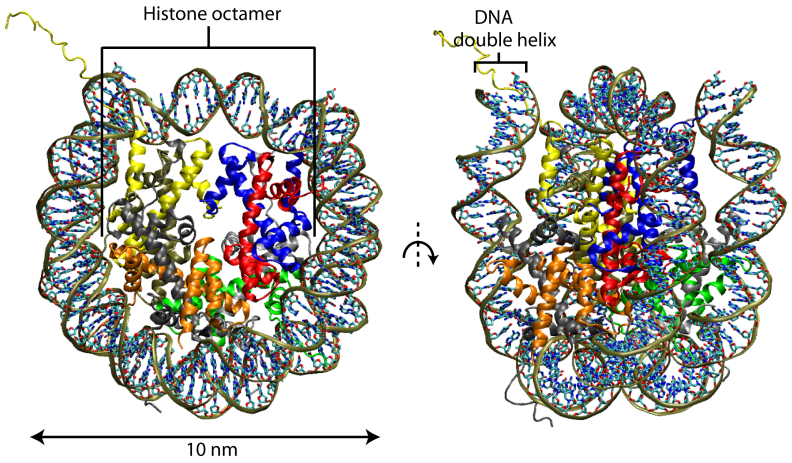


Figure 8: Crystal structure of the nucleosome: complex between the histone core and a 146 bp long DNA fragment [14] (PDB code: 1AOI). You can explore the structure of the nucleosome with the VMD saved state `dna-packing/dna-packing.vmd`.

tone proteins form a disc or "histone core", around which 146 DNA base pairs wrap; this is the fundamental packing unit and is known as the **nucleosome** (Fig. 8). The long DNA chains extend continuously from nucleosome to nucleosome. Each nucleosome is separated from the next by "linker DNA", forming a structure like the one shown in Figure 10a, called "beads in a string". The positions in the DNA sequence where nucleosomes are formed are determined by the flexibility of the DNA in that region. We will look at DNA flexibility in the next section.

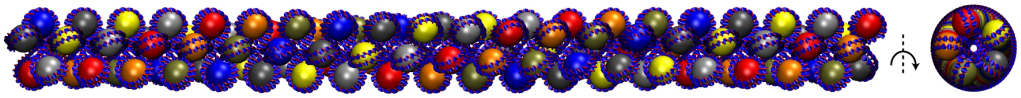


Figure 9: Model of the chromatin fiber. The histone cores are represented by spheres, and the DNA by a twisted ribbon. You can explore the structure of this chromatin model by using the Virtual DNA Viewer plugin in VMD (see Exercise 2) .

Another set of proteins help nucleosomes pack into a higher order structure called **chromatin**. In chromatin, nucleosomes are packed one upon another to generate regular arrays in which DNA is even more condensed. Chromatin is a fiber with a diameter of about 30 nm (Fig. 10b).

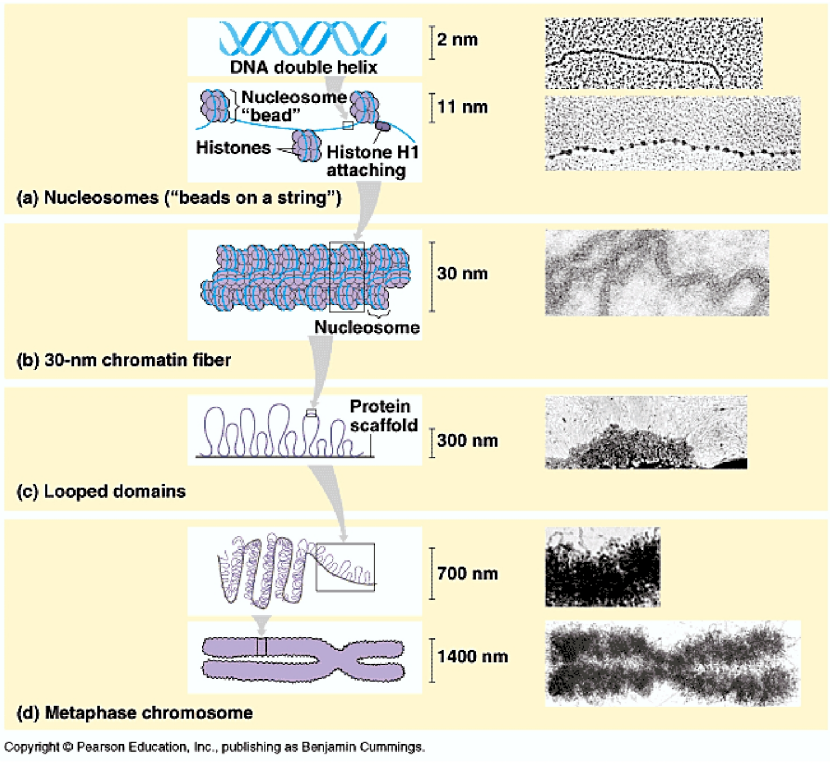


Figure 10: The DNA of a cell is packed in the nucleus in a wide range of length scales. Figure taken from [15].

Even in this fibril form, the DNA would be hundreds of times longer than a cell. Chromatin fibers pack into higher-order structures called **chromosomes**. Despite knowing the structures of chromosomes (Fig. 10, c and d), the underlying mechanism of further compression of chromatin into chromosomes remains unknown. It is known that the shape is important for regulation of gene expression. In humans, there are 23 different chromosomes, and each organism contains two copies of each, one copy inherited from each parent. Twenty-two of the pairs are identical, called homologs. The 23^{rd} chromosome is the sex chromosome; in males, this is a nonhomologous pair.

Exercise 2: Hierarchical packing of DNA. The structure of chromatin depends highly on the ability of DNA to be bent into more compact structures. Load the VMD saved state `dna-packing.vmd` and explore the structure of the DNA molecule presented.

1. Turn off the "DNA" molecule and turn on the "nucleosome" molecule.

   (a) Highlight the positive residues of the protein (you can use the keyword "basic" in a VMD representation) and comment on their location. Note that the phosphate group of each base carries a negative charge, therefore, for 146 base pairs, the DNA around the histone core has a charge of -292e. Find the total charge of the histone core (type the following in the VMD TkConsole: `eval "vecadd [[atomselect 1 protein] get charge]"`). Try to rationalize the difference (Hint: Only one side of DNA is bound to the histone. How does the neutralization of only one side of the DNA increases its ability to bend?).

   (b) Measure the radius of curvature of the DNA wrapped around the histone proteins (hint: assume that the nucleosome is spherical, i.e., constant curvature and measure its radius). You will use this information in a later section when we study the flexibility of DNA.

   (c) Measure the diameter of the whole complex.

   (d) Measure the distance between the ends of the DNA. You will use this information in the next section.

   (e) Approximately, in how many turns does the left-handed DNA superhelix wrap around the histone core? You can obtain this number by looking or by using the radius of curvature and the length of the DNA molecule (146 bp, with 3.4 Å per bp.)

2. Turn off the "nucleosome" molecule. Go to "Extensions", select "Visualization", select "Virtual DNA Viewer". In the plugin window that opens, click on "Default" and select "Chromatin1", then click on "Draw It". This is a coarse-grained model of the chromatin fiber[a]. Explore the chromatin model; note the helical structure of the fiber (Note: This is a graphical object in VMD, you can not pick atoms or measure distances).

   (a) How many histones are there per pitch of the helix?

   (b) Is it a right-handed or a left-handed helix?

   (c) Calculate the approximate diameter of the helix based on the diameter of the complex you obtained.

[a] provided by Tom Bishop

5 DNA flexibility

The DNA double helix in solution has a shape very different from the classical straight DNA configuration in which this molecule is famously depicted. In reality, DNA is a flexible

polymer subject to thermal fluctuations. A polymer is a long chain of similar repeated units, in which each unit is composed of many atoms. At equilibrium, the orientation between the different units is the same, but when subject to thermal fluctuations, the adjacent units differ in direction and are poorly correlated. This results in long molecules of DNA in solution behaving like a random coil (Fig. 11a).

As we saw in the previous section, the flexibility of DNA is essential for fitting long DNA molecules into the cell. Other cellular processes such as genetic regulation and binding of drugs and proteins to DNA also rely on the elastic properties of this molecule. If DNA behaves as a random coil in the cell, it is reasonable to consider using mathematical models of polymers to study its elastic properties.

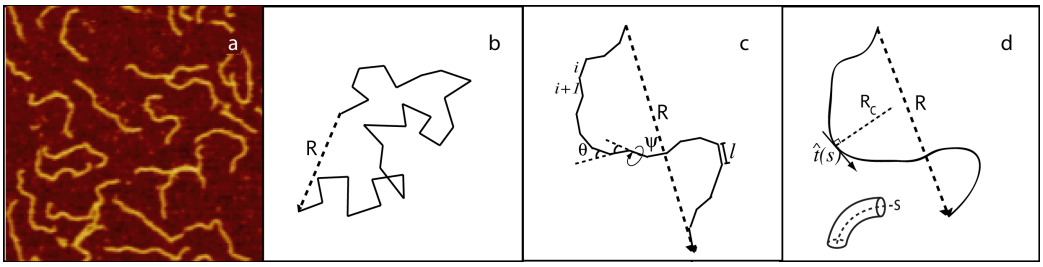


Figure 11: Flexibility of DNA: (a) configuration of a real DNA molecule in solution, taken from [16], (b) freely jointed chain model, (c) discrete worm-like chain model (d) continuous worm-like chain (elastic rod model).

In this section you will see that using a simple model for polymers, DNA can be described with great accuracy. Many models of DNA have been developed that account for various of its physical characteristics, such as its bending and twisting flexibility, electrostatics and sequence-dependent properties [17, 18, 19, 20, 21]. Most of these models introduce corrections to the model presented here, and they reproduce the behaviour of DNA more accurately.

We begin by assuming that DNA is a chain of linked segments of length

l, that are connected by flexible linkers that can rotate freely (Fig. 11b). The **contour length** of the molecule is Nl, where N is the number of segments. The end-to-end distance vector is defined as:

$$\mathbf{R} = \sum_i \mathbf{l}_i, \tag{1}$$

where $\mathbf{l}_i$ is the vector representation of the segment i. The square of the magnitude of $\mathbf{r}$ can be used as a characteristic length of the chain, and is given by

$$R^2 = \left(\sum_i^n \mathbf{l}_i\right)\left(\sum_j^n \mathbf{l}_j\right). \tag{2}$$

Factoring the diagonal terms $i = j$ and taking the average over all possible configurations[1] we get the **mean square end-to-end distance**

$$\left\langle R^2 \right\rangle = Nl^2 + 2\sum_{j>i} \left\langle \mathbf{l}_i \cdot \mathbf{l}_j \right\rangle. \tag{3}$$

For random orientation of the segments (Fig. 11c), the second term on the right hand side is zero, thus

$$\left\langle R^2 \right\rangle = Nl^2, \tag{4}$$

which is the mean square end-to-end distance for the so-called **freely jointed chain** (FJC) model, the simplest model for describing polymers. The FJC model works only for large values of N, and generally describes the conformations of polymers poorly, mainly because it neglects interactions between the polymer with itself (the polymer self-crosses often). As an example, this model gives an average end-to-end displacement $\langle R \rangle = 0$, which is not what you would expect to observe of a polymer in solution.

A widely used measure for the characterization of the configuration of a polymer is the **radius of gyration** (R_G), which measures the root-mean-square distance of the collection of segments from their common center of mass, i.e.

$$R_G^2 = \frac{1}{N+1}\sum_{i=0}^{N} \left(\mathbf{r}_i - \mathbf{r}_{cm}\right)^2 \tag{5}$$

[1] All averages hereupon are taken in the same way.

where $\mathbf{r}_{cm}$ is the vector defining the coordinates of the center of mass ($\mathbf{r}_{cm} = \sum_j m_j \mathbf{r}_j$, m_j and $\mathbf{r}_j$ being the mass and the position of segment j, respectively). The radius of gyration can be measured from particle scattering experiments [22]. The radius of gyration can also be expressed in terms of the average mean square end-to-end distance $\langle R^2 \rangle$. To find this relation, we use the theorem of Lagrange [23]:

$$R_G^2 = \frac{1}{(N+1)^2} \sum_{0 \le i < j \le N} r_{ij}^2 \tag{6}$$

where r_{ij} is the length of the vector connecting segments i and j. Averaging over the statistical ensemble gives

$$\left\langle R_G^2 \right\rangle = \frac{1}{(N+1)^2} \sum_{0 \le i < j \le N} \left\langle r_{ij}^2 \right\rangle . \tag{7}$$

Using (4), for $N = j - i$

$$\left\langle r_{ij}^2 \right\rangle = (j-i) l^2. \tag{8}$$

Thus,

$$\begin{aligned} \left\langle R_G^2 \right\rangle &= \frac{l^2}{(N+1)^2} \sum_{0 \le i < j \le N} (j-i) \\ &= \frac{l^2}{(N+1)^2} \sum_{j=1}^{N} \sum_{k=1}^{j} k \\ &= \frac{l^2}{2(N+1)^2} \sum_{j=1}^{N} (j^2 + j) \end{aligned} \tag{9}$$

where we have introduced the change of variable $k = j - i$. We note that

$$\sum_{j=1}^{N} (j^2 + j) = \frac{N(N+1)(2N+1)}{6} + \frac{N(N+1)}{2}. \tag{10}$$

For large N, this is approximately $N^3/3$, which gives, using (9) and (4),

$$\left\langle R_G^2 \right\rangle \approx \frac{N l^2}{6} = \frac{\langle R^2 \rangle}{6}. \tag{11}$$

Another useful quantity is the **persistence length** (l_p), defined as the average projection of the end-to-end vector onto an arbitrary vector $\mathbf{l}_i$, in the limit of an infinite chain, i.e.

$$l_p = \lim_{N\to\infty} \left\langle \sum_{j=i}^{N} \mathbf{l}_i \cdot \frac{\mathbf{l}_j}{l} \right\rangle . \tag{12}$$

The persistence length is a measure of the rigidity of the molecule, or a measure of the length over which there is significant correlation between the directions of two segments. In order to relate the persistence length to the mean square end-to-end distance, we rewrite (3) in the following way:

$$\begin{aligned} \left\langle R^2 \right\rangle &= Nl^2 + 2\sum_{j>i} \langle \mathbf{l}_i \cdot \mathbf{l}_j \rangle & (13) \\ &= Nl^2 + 2\sum_{j=2}^{N} \langle \mathbf{l}_1 \cdot \mathbf{l}_j \rangle + 2\sum_{j=i+1}^{N}\sum_{i=2}^{N} \langle \mathbf{l}_i \cdot \mathbf{l}_j \rangle & (14) \end{aligned}$$

For large N, the first summation above corresponds to (12), except that it starts from $j = 2$. Therefore,

$$\sum_{j=2}^{N} \langle \mathbf{l}_1 \cdot \mathbf{l}_j \rangle = ll_p - l^2 \tag{15}$$

It is clear that every term in the second summation of (14) also contributes with $ll_p - l^2$. Thus,

$$\left\langle R^2 \right\rangle = Nl^2 + 2N(ll_p - l^2) = 2Nll_p - Nl^2 \tag{16}$$

or

$$l_p = \frac{\langle R^2 \rangle + Nl^2}{2Nl} \tag{17}$$

for large N. As we saw in (4), the mean square end-to-end distance for the FJC model is give by $\langle R^2 \rangle = Nl^2$. Thus, for this model the persistence length is simply given by $l_p = l$.

Now we consider the statistical distribution of the end-to-end vector. Let $\Phi(\mathbf{R}, N)$ be the probability distribution function of the end-to-end vector of

a chain of N segments being $\mathbf{R}$. The conformational distribution of the chain can be considered as a product of independent conformations of the segments

$$\Psi\left(\{\mathbf{r}_n\}\right) = \prod_{n=1}^{N} \psi(\mathbf{r}_n) \tag{18}$$

where $\psi(\mathbf{r}_n)$ is the random distribution of a segment with length l

$$\psi(\mathbf{r}_n) = \frac{1}{4\pi l^2}\delta\left(|\mathbf{r}_n| - l\right) \tag{19}$$

$\Phi\left(\mathbf{R}, N\right)$ is calculated by

$$\Phi\left(\mathbf{R}, N\right) = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \ldots \int d\mathbf{r}_N \delta(\mathbf{R} - \sum_{i=1}^{N} \mathbf{r}_n)\Psi\left(\{\mathbf{r}_n\}\right). \tag{20}$$

Using

$$\delta(\mathbf{r}) = \frac{1}{(2\pi)^3}\int d\mathbf{k} e^{(i\mathbf{k}\cdot\mathbf{r})}, \tag{21}$$

one can rewrite (20) as

$$\begin{aligned}
\Phi\left(\mathbf{R}, N\right) &= \frac{1}{(2\pi)^3}\int d\mathbf{r}_1 \int d\mathbf{r}_2 \ldots \int d\mathbf{r}_N \exp\left(i\mathbf{k}\cdot(\mathbf{R} - \sum_{i=1}^{N} \mathbf{r}_n)\Psi\left(\{\mathbf{r}_n\}\right)\right. \\
&= \frac{1}{(2\pi)^3}\int d\mathbf{k} e^{(i\mathbf{k}\cdot\mathbf{R})} \int d\mathbf{r}_1 \ldots \int d\mathbf{r}_N \prod_{n=1}^{N} e^{(-i\mathbf{k}\cdot\mathbf{r}_n)}\psi(\mathbf{r}_n) \\
&= \frac{1}{(2\pi)^3}\int d\mathbf{k} e^{(i\mathbf{k}\cdot\mathbf{R})} \left[\int d\mathbf{r}_n e^{(-i\mathbf{k}\cdot\mathbf{r}_n)}\psi(\mathbf{r}_n)\right]^N
\end{aligned} \tag{22}$$

The integral over $\mathbf{r}$ is evaluated by introducing polar coordinates (r, θ, ϕ); the reference axis of θ being taken along the vector $\mathbf{k}$, giving

$$\begin{aligned}
&\int d\mathbf{r} e^{(-i\mathbf{k}\cdot\mathbf{r})}\psi(\mathbf{r}) \\
&= \frac{1}{4\pi l^2}\int_0^\infty dr r^2 \int_0^{2\pi} d\phi \int_0^\pi d\theta \sin\theta \exp(-ikr\cos\theta)\,\delta(r-l) \\
&= \frac{\sin kl}{kl}
\end{aligned} \tag{23}$$

where $k = |\mathbf{k}|$. From equations (22) and (23)

$$\Phi(\mathbf{R},\, N) = \frac{1}{(2\pi)^3} \int d\mathbf{k}\, e^{i\mathbf{k}\cdot\mathbf{R}} \left(\frac{\sin kl}{kl}\right)^N . \tag{24}$$

If N is large, $(\sin kl/kl)^N$ becomes very small unless kl is small. For $kl \ll 1$, $(\sin kl/kl)^N$ can be approximated as

$$\left(\frac{\sin kl}{kl}\right)^N \simeq \left(1 - \frac{k^2l^2}{6}\right)^N \simeq \exp\left(-\frac{Nk^2l^2}{6}\right). \tag{25}$$

This approximation holds also for $kl \geq 1$ since both sides of equation (25) are nearly zero in such a case. Thus $\Phi(\mathbf{R},\, N)$ is calculated as

$$\Phi(\mathbf{R}, N) = \frac{1}{(2\pi)^3} \int d\mathbf{k} e^{i\mathbf{k}\cdot\mathbf{R}} \exp\left(-\frac{Nk^2l^2}{6}\right). \tag{26}$$

The integral over $\mathbf{k}$ is a standard Gaussian integral

$$\int_{-\infty}^{\infty} dx \exp(-ax^2 + bx) = (\pi/a)^{1/2} \exp(\frac{b^2}{4a}) \tag{27}$$

If k_α and R_α ($\alpha = x, y, z$) denote the components of the vectors $\mathbf{k}$ and $\mathbf{R}$, then,

$$\begin{aligned}
\Phi(\mathbf{R}, N) &= (2\pi)^{-3} \prod_{\alpha=x,y,z} \left[\int_{-\infty}^{\infty} dk_\alpha \exp(ik_\alpha R_\alpha - Nk_\alpha^2 l^2/6)\right] \\
&= (2\pi)^{-3} \prod_{\alpha=x,y,z} \left(\frac{6\pi}{Nl^2}\right)^{1/2} \exp\left(-\frac{3}{2Nl^2} R_\alpha^2\right) \\
&= (3/2\pi Nl^2)^{3/2} \exp\left(-\frac{3\mathbf{R}^2}{2Nl^2}\right).
\end{aligned} \tag{28}$$

Thus the distribution function of the end-to-end vector is Gaussian.

A correction to the FJC model is made by the **freely rotating chain** or **worm-like chain** (WLC) model. In this model, the n^{th} bond is connected to the $(n-1)^{th}$ bond with a fixed angle θ, but can rotate freely around the $(n-1)^{th}$ bond (Fig. 11c). For this model, the second term in (3) does not average zero. In order to find $\langle \mathbf{l}_i \cdot \mathbf{l}_j \rangle$, we first look at adjacent links

$$\langle \mathbf{l}_i \cdot \mathbf{l}_{i+1} \rangle = l^2 \cos\theta. \tag{29}$$

We can denote $\alpha = \cos\theta$. In order to calculate the average of the projection of bond $i+k$ on i, we note that the inner product of $\mathbf{l}_i$ and projection of $\mathbf{l}_{i+k}$ in the transverse direction of $\mathbf{l}_{i+k-1}$ average zero. The inner product of $\mathbf{l}_i$ and the projection of what remains in the transverse direction of $\mathbf{l}_{i+k-2}$ also average zero, and so on. Thus,

$$\langle \mathbf{l}_i \cdot \mathbf{l}_{i+k} \rangle = l^2 \alpha^k, \tag{30}$$

Introducing this into (3), we have

$$\begin{aligned} \left\langle R^2 \right\rangle &= N l^2 + 2 \sum_{j>i} l^2 \alpha^{j-i} & (31) \\ &= N l^2 + 2 l^2 \sum_{j=1}^{N} \alpha^j \sum_{i=1}^{j-1} \frac{1}{\alpha^i}. & (32) \end{aligned}$$

The sum of the first $n+1$ terms of a geometric series is given by

$$\sum_{k=0}^{n} y^k = \frac{1-y^{n+1}}{1-y}, \tag{33}$$

so that

$$\sum_{i=1}^{j-1} \left(\frac{1}{\alpha}\right)^i = \frac{1-\alpha^{-j}}{1-\alpha^{-1}} - 1 = \frac{1-\alpha^{1-j}}{\alpha-1}, \tag{34}$$

which yields

$$\begin{aligned} \left\langle R^2 \right\rangle &= N l^2 + 2 l^2 \sum_{j=1}^{N} \alpha^j \left(\frac{1-\alpha^{1-j}}{\alpha-1} \right) \\ &= N l^2 + \frac{2 l^2}{\alpha-1} \left[\left(\sum_{j=1}^{N} \alpha^j \right) - N\alpha \right] \end{aligned}$$

Using (33) again

$$\sum_{j=1}^{N} \alpha^j = \frac{1-\alpha^{N+1}}{1-\alpha} = \frac{\alpha-\alpha^{N+1}}{1-\alpha} \tag{35}$$

we obtain

$$\begin{aligned} \left\langle R^2 \right\rangle &= Nl^2 + \frac{2l^2}{\alpha-1}\left[\frac{\alpha-\alpha^{N+1}}{1-\alpha} - N\alpha\right] &(36)\\ &= Nl^2 + \frac{2l^2}{(\alpha-1)^2}\left(N\alpha - N\alpha^2 + \alpha^{N+1} - \alpha\right) &(37)\end{aligned}$$

Rearranging terms leads to

$$\left\langle R^2 \right\rangle = Nl^2\left[\frac{1+\alpha}{1-\alpha} + \frac{2\alpha(1-\alpha^N)}{N(\alpha-1)^2}\right] \tag{38}$$

For large N, we can neglect the last term, obtaining

$$\left\langle R^2 \right\rangle \approx Nl^2\frac{1+\alpha}{1-\alpha}. \tag{39}$$

The mean square radius of gyration for the WLC model can be obtained by combining (38) and (7). As before, we introduce the change of variable $k = j - i$,

$$\left\langle R_G^2 \right\rangle = \frac{l^2}{(N+1)^2}\sum_{j=1}^{N}\sum_{k=1}^{j} k\left[\frac{1+\alpha}{1-\alpha} + \frac{2\alpha}{(\alpha-1)^2}\left(\frac{1}{k} - \frac{\alpha^k}{k}\right)\right] \tag{40}$$

Using (33),

$$\left\langle R_G^2 \right\rangle = \frac{l^2}{(N+1)^2}\sum_{j=1}^{N}\left\{\left(\frac{1+\alpha}{1-\alpha}\right)\frac{j(j+1)}{2} + \frac{2\alpha}{(\alpha-1)^2}\left[j - \frac{1-\alpha^{j+1}}{1-\alpha} + 1\right]\right\} \tag{41}$$

Introducing (10),

$$\begin{aligned} \left\langle R_G^2 \right\rangle &= \frac{l^2}{(N+1)^2}\left\{\left(\frac{1+\alpha}{1-\alpha}\right)\frac{1}{2}\left[\frac{N(N+1)(2N+1)}{6} + \frac{N(N+1)}{2}\right] + \right.\\ &\left. + \frac{2\alpha}{(\alpha-1)^2}\left[\frac{N(N+1)}{2} - \frac{N}{1-\alpha} + N + \sum_{j=1}^{N}\frac{\alpha^{j+1}}{1-\alpha}\right]\right\}\end{aligned}$$

we can evaluate the sum above using (33)

$$\sum_{j=1}^{N} \alpha^{j+1} = \sum_{k=2}^{N} \alpha^{k} = \frac{1-\alpha^{N+1}}{1-\alpha} - 1 - \alpha = \frac{\alpha^2 - \alpha^{N+1}}{1-\alpha}. \tag{42}$$

Thus, rearranging some terms, yields

$$\begin{aligned} \left\langle R_G^2 \right\rangle &= \frac{l^2}{(N+1)^2} \left\{ \left(\frac{1+\alpha}{1-\alpha} \right) \frac{1}{2} \left[\frac{N(N+1)(2N+1)}{6} + \frac{N(N+1)}{2} \right] + \right. \\ &\left. + \frac{\alpha}{(\alpha-1)^3} \left[2(N + \alpha^2 - \alpha^{N+1}) - (1-\alpha)(N^2+N+2) \right] \right\} \end{aligned}$$

For large N, holds then

$$\left\langle R_G^2 \right\rangle \approx \frac{Nl^2}{6} \frac{1+\alpha}{1-\alpha} \approx \frac{\langle R^2 \rangle}{6} \tag{43}$$

The persistence length for the WLC model can be readily calculated from (17) and (39):

$$l_p = \frac{l}{1-\alpha} \tag{44}$$

Exercise 3: Chain of DNA. In this exercise, you will look at a random coil model of a DNA molecule and compare the results given by the WLC model with those measured *in vivo*. Load the VMD saved state `coil.vmd`. This is a WLC representation of DNA. There are 100 configurations. Explore the differences between them. Note that the units of this saved state are nm (generally, VMD saved states are in Å).

1. Find the contour length of the DNA molecule in nm (hint: measure the length of the segment and find the number of segments, given by $N_{atoms} - 1$ for this saved state). How many base pairs does this correspond to?
2. Calculate the persistence length of this molecule (hint: measure the length of the segment and the fixed angle θ between two segments). How many base pairs does this correspond to? Compare to the contour length of the molecule.
3. Measure the end-to-end distance (R) of DNA for each frame and calculate the average. Do this by creating a label between the first and last atoms in the chain and saving the data from the Graphics→Labels menu in VMD.
4. Calculate $\langle R^2 \rangle$ in the same way you calculated $\langle R \rangle$. Compare to the average value for the WLC model for this length of DNA using (39).
5. The radius of gyration of a DNA molecule of this length is 117 nm [22]. Compare with the calculated value using (43). Obtain the average radius of gyration from the sample configurations by sourcing the VMD script `rgyr.tcl`.

Making the length of the segment small, the WLC model can describe a continuous rod, corresponding to the so-called **elastic rod model**. In this case, the rod is described by a coordinate s along its length or centerline (Fig. 11d). Let $\mathbf{r}(s)$ be the position of a point along the chain, and the unitary vector

$$\mathbf{t}(s) = \frac{d\mathbf{r}}{ds} \Big/ \left| \frac{d\mathbf{r}}{ds} \right| \tag{45}$$

the tangent to the chain at this point. The curvature κ is defined as

$$\kappa(s) = \frac{\partial \mathbf{t}}{\partial s}, \tag{46}$$

related to the radius of curvature of the molecule as $\kappa = 1/R_c$ (Fig. 11d).

For small bending, the change in the free energy can be expanded in a Taylor series in κ

$$dU = \frac{dU}{d\kappa}\kappa + \frac{1}{2}\frac{d^2U}{d\kappa^2}\kappa^2 + \dots \tag{47}$$

At equilibrium, the rod is in an energy minimum and thus the first term on the right hand side is zero. Terms of higher order than κ^2 are neglected. Therefore, analogous to Hooke's law, the energy must be of a quadratic form:

$$U = \frac{1}{2}A\int_0^L \kappa^2(s)ds \tag{48}$$

where L is the length of the rod, and $A = d^2U/d\kappa^2$ is a constant related to the rigidity of the molecule.

The bending of the chain is related to thermal motion, and the probability of an energy U due to bending is proportional to the Boltzmann factor $\exp\left(\frac{-U}{k_bT}\right)$:

$$\Psi(\mathbf{t}) \propto \exp\left(-\frac{U}{k_BT}\right) = \exp\left[-\frac{1}{4\lambda}\int_0^L ds\ \kappa^2(s)\right] \tag{49}$$

where

$$\lambda = \frac{k_BT}{2A}. \tag{50}$$

It can be shown that the persistence length is

$$l_p = \frac{1}{2\lambda} = \frac{A}{k_BT} \tag{51}$$

where k_B is the Boltzmann's constant, and T the temperature of the solution. The constant A is called the bending modulus. This quantity has been measured in several experiments; the widely accepted value of A is $1.5 \cdot 10^{-19}\, erg \cdot cm$ [24, 25, 26].

From (49) we can obtain the orientation correlation function of $\mathbf{t}$

$$\langle \mathbf{t}(s) \cdot \mathbf{t}(0) \rangle = \exp\left(-s/l_p\right). \tag{52}$$

Therefore, l_p is the distance measured down the chain at which the orientational correlation function drops to $1/e$ of its value. The persistence length can be measured experimentally [27, 25], and the widely accepted value for the persistence length of B-DNA is 50 nm.

Although the WLC model presented above does not implicitly account for volume exclusion, the introduction of an angular constraint between segments avoids much of the self-crossing, which only occurs between distant segments. A correction can be introduced in case this needs to be accounted for [**?** 23]. The WLC has been a very successful model for describing DNA stretching experiments, and is used for data interpretation very successfully [28].

Exercise 4: Nucleosome revisited. In the previous section, you learned that DNA is wrapped around the histone core in a very compact configuration.

1. How does the persistence length of DNA given in the text compare to the length of the DNA in the nucleosome?
2. How does it compare to the curvature you found in the nucleosome?
3. In exercise 2, you obtained the end-to-end distance of the DNA in the nucleosome. What is the probability that this configuration of DNA occurs in solution (no protein constraints)?
4. According to the WLC model, estimate the energy required to bend DNA into such compact configuration.
5. For this configuration to occur, the interaction energy between the protein and the DNA must be of a similar magnitude. From the structure of the nucleosome you explored before, give examples of interactions between protein and DNA that would account for this.

6 Protein-DNA Interaction

So far, we have considered that a DNA segment shorter than the persistence length behaves as a straight stiff rod. In reality, the sequence of DNA affects its bending and twisting properties. We discussed earlier that nucleosomes form in regions where the DNA is more prone to bend. Furthermore, other environmental factors such as salt concentration, temperature and pressure affect the structure of DNA.

The deviation from the straight double helix at equilibrium can be so prominent that one can observe DNA sequences of the order of the persistence length of DNA that form a circle without being covalently bonded at the ends! (Fig. 12). The study of the sequence-dependent properties of DNA is an active field of research [29, 30, 31], in view of the fact that the sequence

greatly influences the behaviour of the molecule in solution and its interaction with other macromolecules.

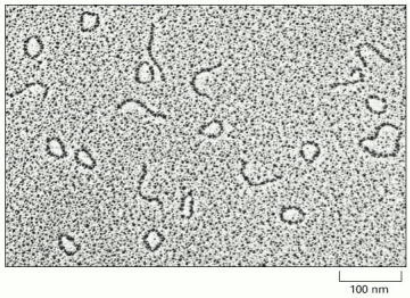


Figure 12: Electron micrograph of DNA segments with 200 base pairs that present highly bent shapes (sometimes circles) due to their sequence. Image taken from [32].

Exercise 5: Intrinsic curvature of DNA. Load the saved state `dna-bend.vmd`. This is an equilibration of DNA with sequence AGGTCACAGTGACCTGGG. Play the equilibration.

1. Note that at the end of the equilibration, the DNA molecule has lost its straight form. Make a figure of the beginning and ending structures, and draw a line along the centerline of DNA to show the bend introduced.

2. Look at the ions that are close to the DNA structure. What kind of ions are they? What is their charge? Can you locate certain regions of the DNA structure in which they stay for long periods of time? Why would ions populate those regions?

We have mentioned that there are several "active" molecular players that manipulate DNA to use the information contained in it. A very important class of molecules that do so are proteins that bind to DNA and manipulate it without altering its double helical structure. A great number of these proteins are involved in gene regulation, such as the *lac* repressor, the CAP protein, and others.

With the whole genome present in the nucleus, the cell has to select which genes are expressed at a given place and time. The selection process is carried out mainly by proteins that recognize sites in DNA, signaling genes to be turned on or off. Several protein motifs have been identified that recognize DNA; before we examine these motifs, let us look at how proteins recognize DNA.

6.1 Proteins have motifs to bind and recognize DNA

In order to recognize DNA, proteins do not require access to the hydrogen bonds between the base pairs in the interior of the double helix to discern between sequences. Instead, they can recognize the edge of a base pair from the outside of the helix, which presents a unique pattern of hydrogen bond donors and acceptors (Fig. 13). In the major groove (c.f. Fig. 5), the distinction between the four base pairs is unique, and therefore proteins generally bind to the major groove.

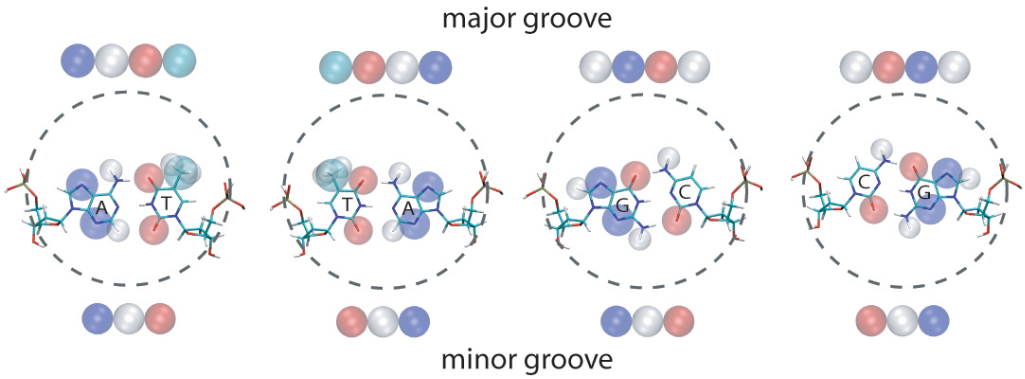


Figure 13: The different base pairs in DNA can be recognized from the helix grooves without the need to open the double helix. The four possible configurations of base pairs are shown. The edge of each base pair contains a distinctive pattern of hydrogen bond donors, acceptors and methyl groups. From the major groove, each base pair projects a unique pattern of features. Adapted from [32]. You can explore the recognition of the base pairs by yourself with the VMD saved state `protein-dna/dna-recognition.vmd`.

Sequence recognition in the grooves at the base pair level shows that proteins can recognize a sequence of DNA very specifically. The physical mechanism for recognition is surface complementarity, where the protein makes a large number of contacts with DNA. Such contacts are of three different kinds: hydrogen bonding, ionic and hydrophobic contacts. The intrinsic bend of DNA for different sequences is known to increase the ability of proteins to differentiate between sequences. Very likely, protein and DNA structures evolved jointly to favor a very tight complementarity. During this evolutionary process, protein DNA-binding motifs emerged. The most common known motifs are explained below and presented in Fig. 14.

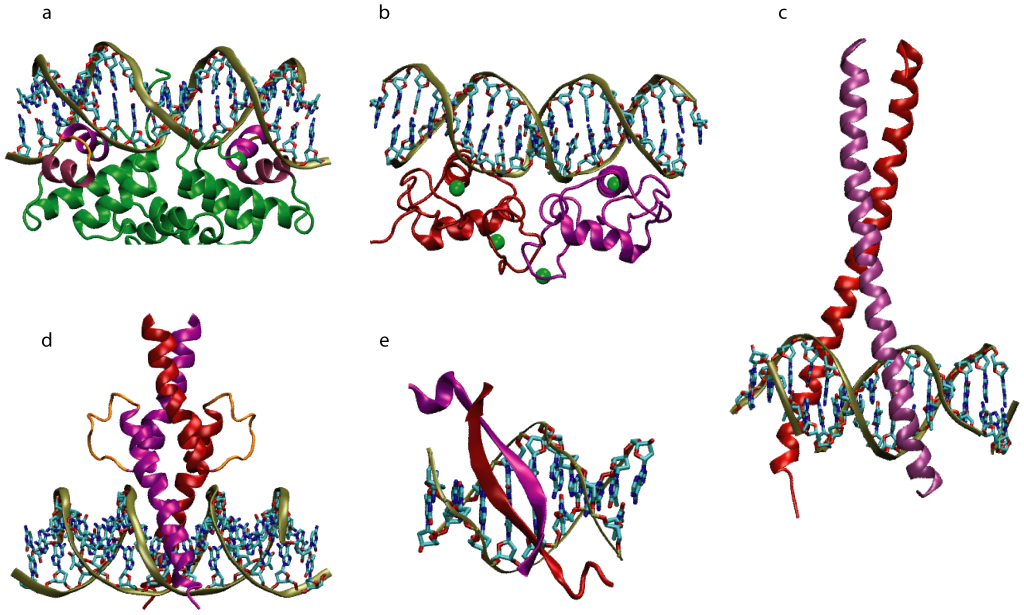


Figure 14: DNA binding motifs: (a) helix-turn-helix (λ-repressor, PDB code 1LMB [33]), (b) zinc fingers (estrogen receptor, PDB code 1GLU [34]), (c) leucine zipper (yeast Gcn4 protein, PDB code 1YSA [35]), (d) helix-loop-helix (transcription factor Max, PDB code 1AN2 [36]), and (e) β-sheet (bacterial *met* repressor protein, PDB code 1CMA [37]). You can explore the protein-DNA binding motifs in these protein complexes with the VMD saved state `protein-dna/binding-motifs.vmd`.

The **helix-turn-helix** consists of two helices of at least six amino acids connected by a short extended strtech of amino acids that form the "turn". One of the helices is called the "recognition helix" as it binds to the major groove of DNA. This motif exists often in a homodimeric form, allowing the protein to bind DNA from both sides of the double helix for best recognition.

The **zinc fingers** exhibit a helix that binds DNA in the major groove, and one or more zinc ions that stabilize the interactions between the protein and DNA. Zinc fingers are often found in nuclear hormone receptors.

The **leucine zipper** consists of two helices joined to form a coiled-coil, held together by interactions between hydrophobic amino acids, often leucines, at each side of the helix. The ends of each helix bind to the major groove of DNA on both sides of the double helix, like a clothespin. This

motif can be a heterodimer, which results in greater specificity.

The **helix-loop-helix**, which is different from the helix-turn-helix motif, consists of a long helix, and a loop that permits a shorter helix to fold back and pack against the first helix, allowing both helices to make contacts with DNA. This motif also exists often in a dimeric form.

Besides helices, proteins can also use **β-sheets** to recognize DNA; they do so by inserting two-stranded β-sheets into the major groove of DNA.

6.2 Nuclear hormone receptors

Nuclear hormone receptors form a class of proteins that get activated by small molecules called hormones, that signal them to bind to specific sequences of DNA and serve as on-off switches for expression of genes in the cell nucleus. These switches control the development and differentiation of skin, bone and behavioral centers in the brain, as well as the continual regulation of reproductive tissues. Hormones are signaling molecules that exert their function by traveling through the bloodstream and interacting with cells in a variety of target tissues.

An example of a nuclear hormone is estrogen. The breast and the uterus, which play central roles in sexual reproduction, are two of the main targets of estrogen. In addition, estrogen molecules act on the brain, bone, liver, and heart. Estrogen acts on these cells by binding estrogen receptors, which are found only in cells that are targets for estrogen action. In this way, estrogen circulates in the bloodstream, but only exerts effects on the cell that contain receptors.

When an estrogen molecule enters a cell that contains receptors, it binds to its receptor, causing it to change its shape. In this new shape, the receptor recognizes a specific sequence of DNA, located near genes that are controlled by estrogen. The binding of the estrogen receptor to DNA induces gene expression, influencing the cell behavior in different ways, depending on the cell type involved. In liver cells, for example, estrogen alters the production of proteins that influence cholesterol levels in the blood.

Exercise 6: Changes in DNA structure induced by a protein. Load the saved state `estrogen-receptor-dna.vmd`. The molecule you loaded is part of the human nuclear estrogen receptor protein [34]. This is an equilibrated structure of the estrogen receptor-DNA complex. The DNA sequence it recognizes is the same as the one you looked at in exercise 5. Look at the different representations of the molecule (turn on the ones that are not shown).

1. Make a figure of the DNA similar to the one you made for exercise 5, and draw a line through the centerline of DNA. Comment on the change of shape introduced into the DNA by the protein.

2. Examine the protein-DNA interface. Can you recognize any of the protein-DNA interaction motifs mentioned above? Make a figure showing the main features of the motif.

3. Modify the figure you just created to show some of the amino acids that make contact with DNA. What are the main interactions (hydrophobic, ionic, hydrogen bonds) that govern the protein-DNA recognition in this case?

4. The zinc ions serve to stabilize the structure of the protein, and sometimes make contacts with the DNA phosphate. Make a figure showing what kind of residues surround the zinc ions.

5. Look for other ions close to the surface of DNA. Note that the ions you saw in exercise 5 have been replaced by protein contacts. Are there any charged amino acids compensating for this replacement? Explain why.

References

[1] W. Saenger, editor. *Principles of Nucleic Acid Structure*. Springer–Verlag, New York, NY, 1984.

[2] F. Miescher. Ueber die chemische Zusammensetzung der Eiterzellen. *Med.-Chem. Unters.*, 4:441–460, 1871.

[3] O. T. Avery, C. M. MacLeod, and M. McCarty. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79:137–158, 1944.

[4] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*. Worth Publishers, 3rd edition, 2000.

[5] A. D. Hershey and M. Chase. Independent functions of viral proteins and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.*, 36:39–56, 1952.

[6] E. Chargaff. Structure and function of nucleic acids as cell constituents. *Fed. Proc.*, 10:654–659, 1951.

[7] L. Pauling and R. B. Corey. A proposed structure for the nucleic acids. *Proc. Natl. Acad. Sci. USA*, 39:84–97, 1953.

[8] J. D. Watson and F. H. C. Crick. A structure for deoxyribose nucleic acids. *Nature*, 171:737–738, 1953.

[9] B Maddox. The double helix and the 'wronged heroine'. *Nature*, 421:407–408, 2003.

[10] R. E. Franklin and R.G. Gosling. The structure of sodium thymonucleate fibers. I. the influence of water content. *Acta Cryst.*, 6:673–677, 1953.

[11] A. H. Wang, G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel, and A. Rich. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, 282:680–686, 1979.

[12] A. Rich and S. Zhang. Timeline: Z-DNA: the long road to biological function. *Nat. Rev. Genet.*, 4:566–572, 2003.

[13] J. Wayne, T. Siler, A. Suzanne, K. Clarke, A. Denes, M. J. Grey, R. Kavenoff, T. Kovachevich, D. Kremers, J. Newman, and M. R Rich. Models, metaphors and matter: Artists and scientists visualize scientific concepts. *"Art Journal"*, 55(1):33–43, 1996.

[14] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389:251–260, 1997.

[15] N. Campbell. *Biology.* Benjamin Cummings, 6th edition, 2001.

[16] M. Lysetska, A. Knoll, D. Boehringer, T. Hey, G. Krauss, and G. Krausch. UV light-damaged DNA and its interaction with human replication protein A: an atomic force microscopy study. *NAR*, 30:2686–2691, 2002.

[17] Alexander Balaeff, L. Mahadevan, and Klaus Schulten. Modeling DNA loops using the theory of elasticity. *Phys. Rev. E*, 73:031919, 2006. (23 pages).

[18] W. K. Olson and V. B. Zhurkin. Modeling DNA deformations. *Curr. Opin. Struct. Biol.*, 10:286–297, 2000.

[19] J. F. Marko. DNA under high tension: Overstretching, undertwisting, and relaxation dynamics. *Phys. Rev. E*, 57(2):2134–2149, 1998.

[20] A. V. Vologodskii and N. R. Cozzarelli. Conformational and thermodynamic properties of supercoiled DNA. *Annu. Rev. Biophys. Biomol. Struct.*, 23:609–643, 1994.

[21] T. Schlick. Modeling superhelical DNA: recent analytical and dynamic approaches. *Curr. Opin. Struct. Biol.*, 5:245–262, 1995.

[22] P. Nelson. *Biological Physics*. W.H. Freeman and Company, New York, 2004.

[23] P. J. Flory. *Statistical Mechanics of Polymer Chains*. Hanser Publishers, Cincinnati, 2003.

[24] T. R. Strick, J.-F. Allemand, D. Bensimon, and V. Croquette. Stress-induced structural transitions in DNA and proteins. *Annu. Rev. Biophys. Biomol. Struct.*, 29:523–543, 2000.

[25] P. J. Hagerman. Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.*, 17:265–286, 1988.

[26] P. J. Heath, J. B. Clendenning, B. S. Fujimoto, and J. M. Schurr. Effect of bending strain on the torsion elastic constant of DNA. *J. Mol. Biol.*, 260:718–730, 1996.

[27] S. G. Baumann, S. B. Smith, V. A. Bloomfield, and Carlos Bustamante. Ionic effects on the elasticity of single DNA molecules. *Proc. Natl. Acad. Sci. USA*, 94(3):6185–6190, 1997.

[28] C. Bustamante, Z. Bryant, and S. B. Smith. Ten year of tension: single-molecule DNA mechanics. *Nature*, 421:423–427, 2003.

[29] Y. Yang, T. P. Westcott, S. C. Pedersen, I. Tobias, and W. K. Olson. Effects of localized bending on DNA supercoiling. *Trends Biochem. Sci.*, 20(8):313–319, 1995.

[30] B. D. Coleman, W. W. Olson, and D. Swigon. Theory of sequence-dependent dna elasticity. *J. Chem. Phys.*, 118(15):7127–7140, 2003.

[31] Wilma K. Olson, Andrey A. Gorin, Xiang-Jun Lu, Lynette M. Hock, and Victor B. Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, 95:11163–11168, September 1998.

[32] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of The Cell*. Garland Science, New York & London, 4th edition, 2002.

[33] L. J. Beamer and C. O. Pabo. Refined 1.8 Å crystal structure of the a repressor-operator complex. *J. Mol. Biol.*, 227:177–196, 1992.

[34] B.F. Luisi et al. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*, 352:497–505, 1991.

[35] T. E. Ellenberger and C. J. Brandl and K. Struhl and S. C. Harrison. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell*, 71:1223–1237, 1992.

[36] A. R. Ferre-D'Amare and G. C. Prendergast and E. B. Ziff and S. K. Burley. Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature*, 363:38–45, 1993.

[37] W. S. Somers and S. E. V. Phillips. Crystal structure of the met repressor-operator complex at 2.8 Åresolution reveals DNA recognition by beta-strands. *Nature*, 359:387–393, 1992.