1

Structural and Functional Regulation of DNA: Geometry, Topology and Methylation

C. Auclair

The work of Rosalind Franklin, then Watson and Crick [1], established the architecture of deoxyribose nucleic acid (DNA), carrier of all genetic information. The idea that DNA was structurally organised in the form of a double helix comprising two antiparallel and complementary polymer chains was one of the great scientific discoveries of the twentieth century. It revealed not only the way in which genetic information is stored, but also the mechanism by which the genetic code is read, and the way this code can be faultlessly copied from one cell to another during cell division.

The structural organisation of genomic DNA varies significantly from one organism to another, or from one cell to another, depending as it does on the physiological constraints specific to each organism or tissue. This complexity can be observed in particular in the diversity of genomic sequences, the size of the human genome being something like 3 gigabases for about 30,000 genes, whereas yeast, a lower eukaryotic organism, only possesses 6,200 genes for a size of 13 megabases (see Table 1.1).

The fraction of protein-coding sequences is also highly variable (1.4% for the human genome, 68% for the yeast genome), and so too is the size of the genes. Particularly interesting is the variation in the content of G+C bases, which determines the overall stability of the DNA helices. Sequences rich in G+C bases are involved in the key processes regulating gene expression and probably in a dominant way in dynamical processes. An important point is the possibility of methylating cytosines, especially the CpG sequences, a crucial process in the control of gene expression. The presence of alternating sequences of GC base pairs, associated with the methylation of the cytosines in these sequences, favours in particular the transition from the B to the Z conformation (see below). Within a given genome, the G+C content can vary significantly, reaching 80% in some regions of mammal genomes, and there seems to be a correlation between the GC base content (especially GCs3) and the gene density in the relevant region.

The complexity of DNA depends directly on the kinds of sequences, but is also characterised by the broad range of micro- and macrostructures resulting

P. Boisseau et al. (eds.), *Nanoscience*,
DOI: 10.1007/978-3-540-88633-4_1,

Table 1.1. Genomic characteristics of some eukaryotic organisms

(*S. cerevisiae*)	Yeast (*C. elegans*)	Nematode (*D. melanogaster*)	Drosophila	Human
Size (Mb)	13	100	180	3,000
[G+C] content	38%	36%	43%	41%
Number of genes	6,200	19,100	13,600	∼30,000
Coding fraction	68%	27%	13%	1.4%
Number of exons per gene	1.04	5.5	4.6	8.7
Size of genes (kb)	1.4	2.7	3	28

from the physicochemical constraints imposed by sequencing and pairing of bases. The wide range of possible conformations of DNA, not to mention molecular arrangements such as cruciform, triple helix, and base tetrad structures, among others, plays a key role in the way genomes work, especially through their specific recognition by proteins carrying out important genetic functions.

One could describe DNA and its protein environment as a nanoworld of the most complex kind. Naturally, this complexity reflects the functions the system has to fulfill. However, one can nevertheless identify certain representative elements in the workings of the genetic machinery. In this context, the aim of the present chapter will be to provide, with the help of some examples, a succinct review of certain features arising from the geometric and topological flexibility of DNA, and to describe rather briefly the structural modifications related to methylation of cytosines and the functional modifications that result from them.

1.1 Geometry of the DNA Double Helix

The conformation originally described by Watson and Crick was a double helix known as the B conformation (see Fig. 1.1), which is the one must often observed in the natural state.

However, the double helix can occur in three geometric forms denoted A, B and Z, characterised in particular by different degrees of hydration. These various conformations of the DNA are made possible by an extraordinary level of geometric freedom allowed between the constituents of DNA. The main point is the existence of conformers: orientation of the sugar constituent (C2′-endo or C3′-endo) and orientation of the base with respect to the sugar (syn or anti) (see Fig. 1.2).

The DNA helix is characterised by the C2′endo/anti conformation in the case of the B conformation, and the C3′endo/anti conformation for the A conformation. The situation is a little more involved in the case of the Z

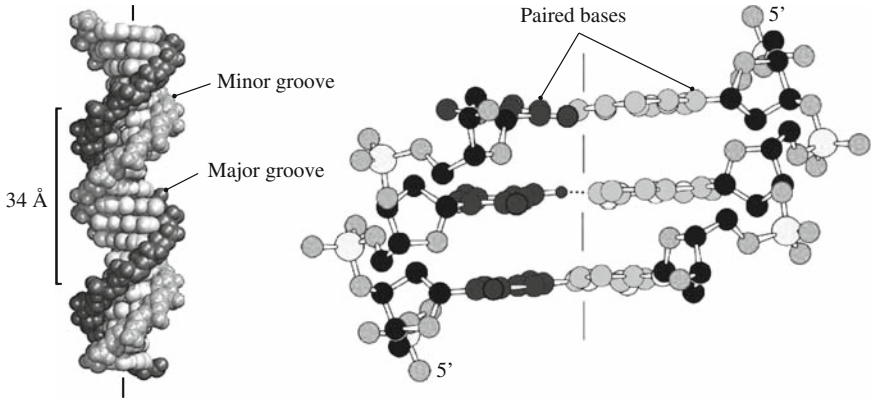

Fig. 1.1. *Left*: Molecular model of a DNA double helix in the B conformation. *Right*: Detail of base pairing

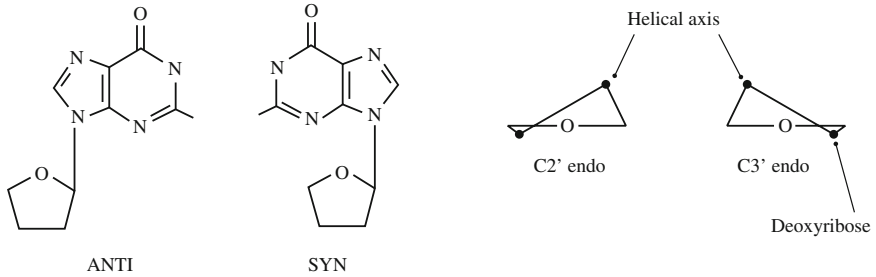

Fig. 1.2. Orientation of the base (guanine) with respect to the deoxyribose (anti and syn) and orientation of the deoxyribose (C2′ endo and C3′ endo)

conformation, where the purine bases adopt the C3′endo/syn conformation while the pyrimidine bases adopt the C2′endo/anti conformation. In fact, the structural characteristics and the evolution of the DNA double helix toward one of the possible conformations are conditioned by the parameters known as twist, roll, and tilt specifying the helix and the set of six torsion angles α, β, γ, δ, ε, and ζ of the phosphate–sugar backbone.

Figure 1.3 shows the main geometric arrangements that the stacks of bases can adopt in relation to one another. The twist is the angle of rotation of two adjacent stacks of base pairs about the helical axis of symmetry. The roll corresponds to the angle of rotation of two adjacent stacks of base pairs about the third axis of the helix. The tilt is the angle of rotation of two adjacent stacks of base pairs about the pseudosymmetry axis of the helix. It should be noted that the position of the plane of each paired base can also vary relative to the other base in the pair (opening, propeller, and buckle), thus increasing the flexibility of the whole construction.

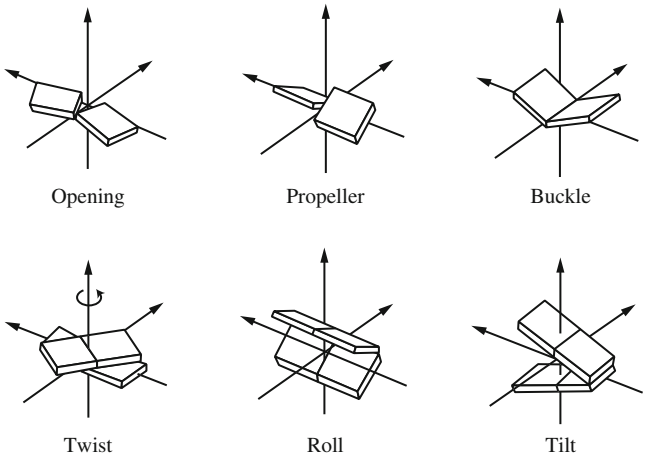

Fig. 1.3. Helical parameters specifying the geometry of the DNA helix. *Arrows* indicate the axes of symmetry of the helix

Table 1.2. Estimated values of the parameters specifying a B-DNA helix for different sequences. Angles are given in degrees and correspond to the angles between a stack of base pairs and its nearest neighbour. The distance between stacks of base pairs is set to 3.4 Å. The method here is due to Bolshoy [3]

Sequence	Twist	Roll	Tilt
A	–	–	–
A	35.67	−6.5	3.2
A	35.67	−6.5	3.2
T	–	–	–
A	36	0.9	0
T	31.2	2.6	0
G	–	–	–
G	33.67	1.2	−1.8
G	33.67	1.2	−1.8
C	–	–	–
G	29.8	6.7	0
C	40.1	−5	0

There are rules [2,3] for assessing the static structure of the DNA molecule as a function of its sequence. The problem is to estimate the above helical parameters. Using the computation program devised by the Georgia Institute of Technology [4], it is easy to demonstrate the dependence of the sequence on the these parameters. As an illustration, Table 1.2 shows the changes in the twist, roll, and tilt for two non-alternating sequences, viz., −AAAA... and −GGGG..., and for two alternating sequences, −TATA... and −CGCG....

Among the remarkable points regarding the data in the table, note the high twist and roll engendered by poly A sequences. This largely explains the curvature of the helix observed in regions with this type of sequence. Indeed, such non-alternating sequences can induce significant deviations from helical symmetry. Modelling and molecular dynamics simulations have shown that this kind of planar curvature results from a tendency to stretch the sugar–phosphate backbone, causing compression with a modification of the torsion and the alignment of the bases, which in turn leads to curvature. This kind of curvature can play an important part in interactions between DNA and its ligands, especially ligands of the minor groove and multimer proteins.

Furthermore, high-resolution structural analysis of the torsion angles of the B helix (angles e C3′–O3′ and z O3′-P) has revealed the existence of BI and BII subconformations [5]. Note also that the ionic strength of the environment of the helix and its level of hydration play predominant roles in determining the conformation that is finally adopted.

Apart from its ability to store and transmit genetic information, it has now been shown that the DNA molecule can itself carry out some degree of regulation, controlling among other things the reading of the code, i.e., the level of gene expression depending on the specific needs of each cell. This regulatory ability arises partly from the presence of regulating sequences located upstream of gene reading frames, and partly from the many different structural and conformational modifications to which the DNA molecule may be subjected. Structural patterns arising from the kind of sequences and/or the architectural organisation of DNA are recognised by a wide range of different proteins, which then act as effectors for genetic functions. This system is effectively based on a kind of molecular recognition, suggesting a great diversity and high level of flexibility in the regulating structural patterns. In the light of more recent work, the phenotype of a cell, i.e., its functional characteristics and morphology, can also be considered to depend just as much on the kinds of genes as on the architectural organisation of the genome. From this standpoint, the available conformational and topological variants of DNA look more and more like key regulatory parameters.

1.2 The Z Conformation of DNA

Since the initial description of the structural parameters specifying the DNA double helix in its most common form (the B conformation depicted in Fig. 1.1), a great deal of further work has shown that this architectural organisation exhibits an exceptional level of flexibility, able to generate an enormous number of variants, each of which would appear to contribute in a crucial way to the functional activities of DNA, including regulating the expression of coded genes.

A representative example of the conformational flexibility of DNA is provided by the Z conformation (see Fig. 1.4). Indeed, it is a striking fact that,

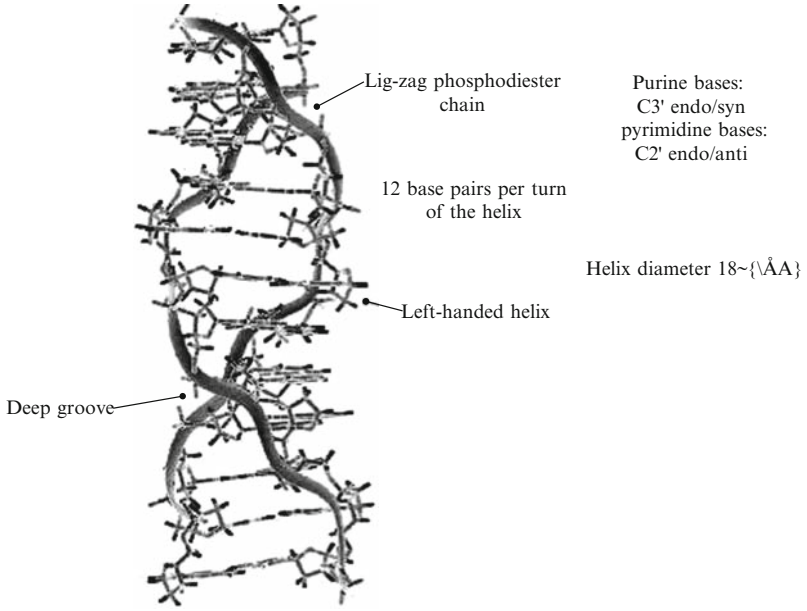

Fig. 1.4. Molecular model of a DNA double helix in the Z conformation, the left-handed helix generally favoured in regions rich in G-C base pairs

following a relatively small energy input, the DNA double helix can change from its B conformation (right-handed helix) to a Z conformation, a left-handed helix with very different structural characteristics (see Table 1.3). In fact, the activation energy required for transition from the B conformation to the Z conformation is about $22\,\mathrm{kcal\,mol^{-1}}$, roughly equivalent to the energy needed for base pair breaking, whether the DNA is in the B or the Z form [6]. This indicates the facility with which the transition can occur from an energetic point of view and also the identity of the molecular dynamics of the two conformations.

In its B conformation, the angle of rotation between consecutive bases is about 36°, so that there are 10 pairs of bases for each turn of the helix. One turn occupies about 34 Å, implying a distance of 3.4 Å between two consecutive base pairs. In its Z conformation, the DNA helix is characterised by a high value of the distance occupied by one turn of the helix and the presence of 12 base pairs per turn. One also finds a reduction in the diameter of the helix and only one rather deeply indented groove. The Z helix is also characterised by the coexistence of glycosidic bonds in the syn form for purine bases and in the anti form for pyrimidine bases. One consequence of this bond alternation is that the sugar–phosphate backbone of the helix adopts a zig-zag shape rather than a regular spiral as it does in B-type DNA. Another consequence that is probably important from a functional point of view is the non-uniform

Table 1.3. Comparative structural parameters for the B and Z forms of the DNA double helix

Structural parameters	B	Z
Orientation of helix	Right-handed	Left-handed
Repetition	1 bp	2 bp
Rotation/bp	35.9°	−30°
Average number of bp/turn	10.0	12
Angle between base and axis	−1.2°	−9°
Distance between bp along axis	3.4 Å	3.7 Å
Distance for one turn of the helix	34 Å	45 Å
Average torsion	+16°	0°
Glycosidic bond	anti	C: anti, G: syn
Sugar conformation	C2′-endo	C: C2′-endo, G: C2′-exo
Diameter of helix	20 Å	18 Å

distribution of negative charges along the helix, which can significantly modify interactions with the various ligands, especially protein ligands.

The transition from the B conformation to the Z conformation is favoured by a high ionic strength, indicating a strong electrostatic component in the stability of the two conformers. The compacted form of the Z helix is to a large extent stabilised by a high salt concentration and the particular hydration network solvating the charges and polar groups [7].

There is a connection between the nucleotide sequence in the DNA and the ability of the double helix to adopt a Z conformation. This is particularly so for alternating sequences of type GCGCGC. The determining factor explaining this phenomenon is in fact the presence of an amine in position 2 of the guanine, which stabilises the Z helix. Replacing the guanine by inosine (which is a guanine but without the amine in position 2) considerably destabilises the Z conformation [8]. In a poly d(G-C) alternating sequence, the difference of free energy between the B form and the Z form is slight, being about $0.33\,\mathrm{kcal\,mol^{-1}}$ per base pair [9], and this favours the great stability of the Z conformation.

The tendency of pyrimidine–purine dinucleotides to form sequences of Z-type DNA is, in decreasing order:

$$\mathrm{m5CG} > \mathrm{CG} > \mathrm{TG} = \mathrm{CA} > \mathrm{TA} \qquad \text{(m5C:5-methylated cytosine).}$$

This brings us to the interesting role played by methylation of cytosines, which also tends to stabilise the Z conformation. This stabilising effect probably comes from the fact that the methyl group at position 5 on cytosine prevents the setting up of a hydration network which stabilises the B form. In fact, alternating sequences of G-C type can occur in the B form or in the Z form and are subject to changes in equilibrium with the environment of the helix, energy constraints applied to the helix, and the presence of chemical or protein

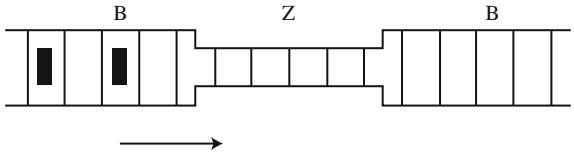

Fig. 1.5. Schematic view of a DNA sequence in the Z conformation inserted into a B helix. Intercalating a plane molecule in the B sequences adjacent to the Z-type sequence induces a Z-to-B transition at a distance

ligands. Intercalating agents bound to sequences adjacent to Z-conformation regions (see Fig. 1.5) can easily induce the Z-to-B transition, whereas groove ligands have almost no effect [10].

This way of inducing the transition results from the significant unwinding of the B helix, which induces stresses from a distance. On the other hand, torsion stresses induced by the transcription process clearly favour the transition from B to Z. This is consistent with the fact that the Z conformation is favoured by negative superhelicity in the DNA helix (see Sect. 1.3). From the biological point of view, note that regions close to transcription initiation sites are rich in sequences favouring the B-to-Z transition. Furthermore, as the RNA polymerase moves along the transcribed DNA strand, regions of increasingly negative superhelicity form upstream. Globally, the formation of Z helices near promoter sequences seems to stimulate transcription. It seems likely that the possibility of conformational transitions in regions of high topological stress provide a way of minimising the energy required for the process to go ahead, or even a way of temporarily stabilising favourable architectural arrangements.

From a structural point of view, the coexistence of B and Z forms raises many questions, and the architectural arrangement of the helix at the B–Z junction always seemed somewhat mysterious, until teams led by A. Rich and K.K. Kim succeeded in resolving this junction at 2.6 Å using X-ray diffraction [11].

The structure of the B–Z junction is characterised by pair breaking in an A-T base pair, while the unpaired bases are extruded outside the helix as shown in Fig. 1.6. This pair breaking corresponds to an energy relaxation which allows the system to maintain a regular stacking of the bases. This is consistent with the free energy estimate of $5\,\mathrm{kcal\,mol^{-1}}$ for the B–Z junction as put forward by Peck and Wang.

There can be no doubt that the special structure of this junction characterised by the two extruded bases could constitute a motif for recognition by certain proteins.

Several proteins have been identified that look likely to bind preferentially to Z-type DNA. This is the case for the protein E3L of the vaccinia virus, the protein AF2008 of *Archaeoglobus fulgidus*, the protein DLM-1, and the RNA-editing enzyme ADAR-1. These proteins have a Z-DNA binding region

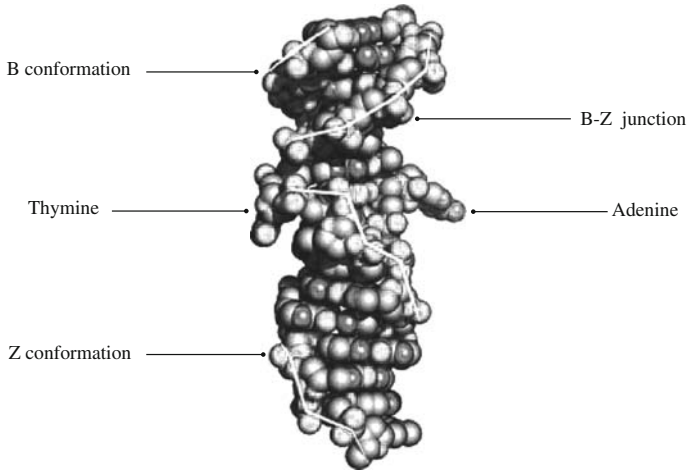

Fig. 1.6. Structural organisation of the B–Z junction [11]

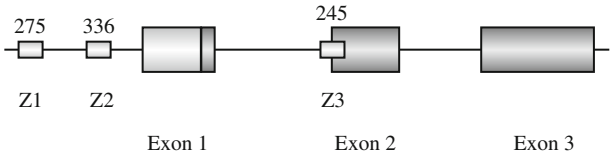

Fig. 1.7. Region of the c-myc gene in which there are three sequences Z1, Z2, and Z3 assuming the Z conformation during transcription [14]

(Za) located in the N-terminal region. The 3D structure of the Za region of ADAR-1 shows that there is a helix–turn–helix motif and a single bond with a guanine in the syn conformation, characteristic of Z-DNA [12].

The 3D structure of the protein AF2008 from *Archaeoglobus fulgidus* resolved at 1.55 Å [13] reveals the dimeric organisation of the protein, in which the Z-DNA binding region of each monomer is separated by 45 Å, a distance corresponding to one turn of the DNA helix in the Z conformation.

Many observations have now been made suggesting that the Z conformation of DNA plays a determining role in the control of gene expression.

Using permeabilised nuclei of U937 cells, it has been shown that, during transcription of the c-myc gene, three Alu I restriction sequences located in the vicinity of the promoter sequences adopted a Z conformation (see Fig. 1.7) [14].

At the end of transcription, the Z conformations rapidly disappear, this being related to relaxation of negative supercoiling by topoisomerase I. Many other examples can be found in the literature. One recent observation has confirmed the crucial biological role of Z-DNA. Indeed, it has been shown that integrity of the Z-DNA binding region of the protein EL3 is crucial for the virulence of the vaccinia virus [15]. Deleting the 83 amino acids of the N-terminal region of EL3 totally deactivates the virus. However, replacing these

83 amino acids by the Z-DNA binding regions of DLM1 or ADAR1 does not affect the virulence of this virus.

The early detection of Z helices in vivo by virtue of the fact that it is easy to obtain anti Z antibodies long remained a mystery. However, it is clear today that the B-to-Z transition has an important regulatory effect with regard to genetic functions. Stresses induced when the negative superhelicity increases, which is almost the rule during replication and transcription of DNA, systematically lead to the formation of Z helices in poly-purine/poly-pyrimidine sequences. This is also the case in nucleosomal DNA (see Chap. 2), especially in the close-packed regions of chromatin. Proteins binding on the Z helices can thereby act to stabilise the structure and possibly also to activate or repress transcription.

The biological significance of the Z form of DNA has been indirectly confirmed by the presence of a large number of sequences for which the probability of undergoing a B-to-Z transition is high, and this for a moderately high level of superhelicity. As an example, on the human chromosome 22, there are 7,580 regions exhibiting these characteristics [16].

1.3 Supercoiled DNA

A second example demonstrating the molecular diversity of DNA helices is provided by the topological variations of DNA so well characterised in circular DNA. In bacteria, some viruses, and mitochondria, DNA helices do indeed adopt a closed circular form as shown in Fig. 1.8.

Circular DNA helices can arrange themselves in space to form positive or negative supercoils, leading to what are known as topoisomers. The

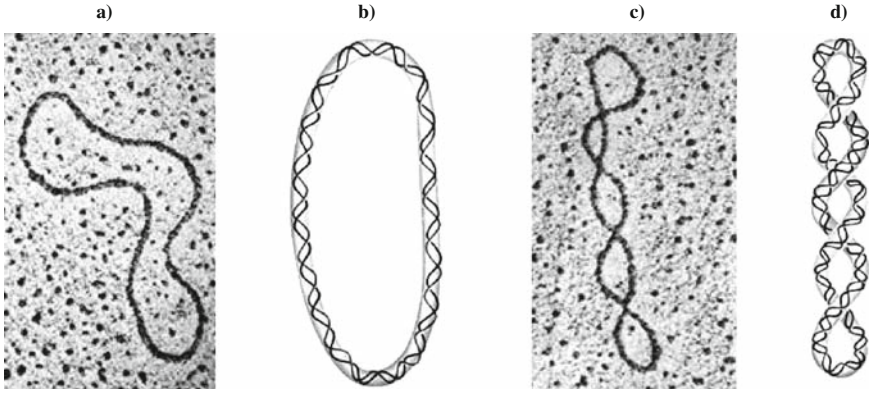

Fig. 1.8. (**a**) and (**b**): Relaxed circular DNA. (**c**) and (**d**): Supertwisted DNA with negative supercoiling. (**a**) and (**c**) are electron microscope images, while (**b**) and (**d**) are diagrammatic

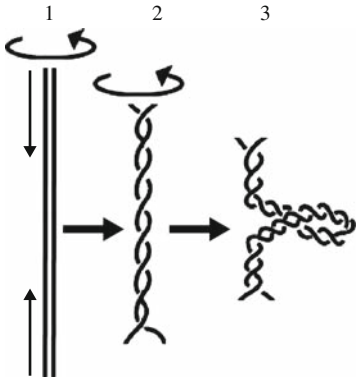

Fig. 1.9. Writhe of a straight, double-stranded helix leading to supercoiling. (1) $T = 0$, $W = 0$. (2) $T = 8$, $W = 0$, $L = 8$. (3) $W = 1$, $\Delta L = 1$

supercoiling of DNA is characterised by three parameters: the linking number L, the twist T, and the writhe W. These are related by the simple formula

$$L = T + W \ .$$

In a relaxed circular DNA, the writhe corresponding to the degree of supercoiling is zero. In supercoiled DNA (see Fig. 1.8c), $W = -4$, and the linking number is less than the twist. Figure 1.9 illustrates schematically the supercoiling phenomenon and the relation between twist and writhe.

The difference in supercoiling energy between two topoisomers depends on the square of the change in linking number:

$$\Delta G_{\mathrm{sc}} = \frac{1}{N} KRT(\Delta L)^2$$

In the above example, $(\Delta L)^2 = 16$.

It is a striking thing that DNA always has negative supercoiling in bacteria, but positive supercoiling in archaeobacteria, and both geometries in eukaryotes.

The degree of supercoiling and its orientation depend on the angle of rotation between consecutive bases, itself determined by the distance between base pairs. As an example, intercalating a planar molecule such as ethidium bromide, or antitumor molecules such as the ellipticines and acridines, between base pairs of a circular DNA induces a rotation of adjacent base pairs of the order of $-20°$ to $-26°$ at the intercalation site. Depending on the increasing number of intercalating molecules, this rotation induces a relaxation of negatively supercoiled DNA first to circular DNA, then to positively supercoiled DNA. However, ligands binding in the minor groove of DNA, such as netropsin, induce a twist of a few degrees, favouring the formation of a negative superhelicity and eventually a close-packing of the DNA. Non-closed DNA can also undergo negative or positive supercoiling. This happens, in particular, for nucleosomal DNA in chromatin structures (see Fig. 1.10).

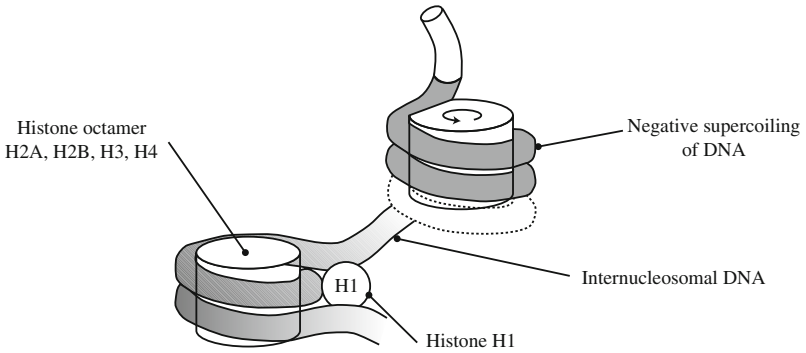

Fig. 1.10. Schematic view of nucleosomal DNA. The DNA molecule is wrapped around a histone octamer stabilised by histone H1

In nucleosomes, the wrapping of DNA around histones corresponds to negative superhelicity. Note that, excepting the nucleosome case, DNA can adopt a positive superhelicity. The situation is in fact relatively complex insofar as the nucleosome can fluctuate between three conformational states. Two of these conformations are characterised by a wrapping of about 1.7 turns around the octamer and a negative or positive crossing over of the leading and lagging DNA strands. The third conformation corresponds to a more open architecture, characterised by a partial wrapping of about 1.45 turns without crossing over of the leading and lagging DNA strands, following an unwrapping of the DNA on either edge of the nucleosome. The possibility of attaining these conformations is regulated by the wrapped DNA sequence. This affects the local twist of the helix, which in its turn acts on the organisation of the leading and lagging DNA strands. It is clear that the nucleosomal architecture must have a very significant level of conformational flexibility. This is illustrated by the fact that histone (H3–H4)2 tetramer can associate equally well with either a positively or a negatively supercoiled DNA minicircle [17].

The extent of DNA supercoiling is also directly regulated by enzymes called topoisomerases which play a determining role in many genetic functions such as replication, transcription, and so on. These enzymes act by cutting then religating DNA strands, one strand for type I topoisomerases and two for type II topoisomerases. In prokaryotic organisms, topoisomerase I reduces negative supercoiling, while gyrase, acting as a type II topoisomerase, preferentially reduces positive supercoiling. Changes in structure and close-packing introduced in this way play a determining role in regulating DNA functions. Moreover, the enzymes known as helicases can unwind the DNA helix in order for replication to take place and can in this way induce topological changes in the vicinity of the replication fork. The same is true during transcription, where the RNA polymerase itself has a helicase action. As an example, the RNA polymerase in *E. coli* unwinds the DNA helix by 140° [18]. The key observations in this field were made by Liu and Wang [19], who showed that, during

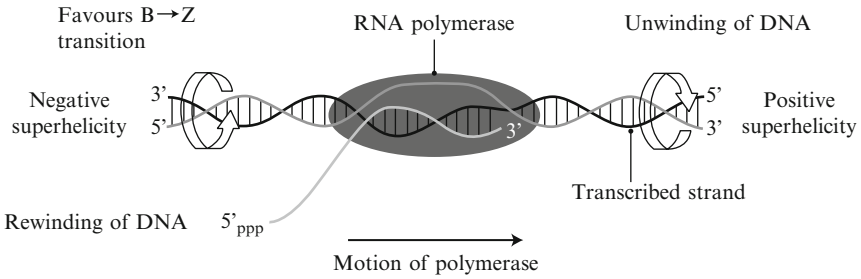

Fig. 1.11. Winding and unwinding of DNA during transcription

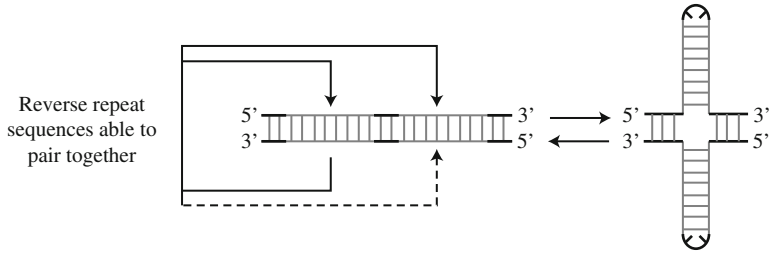

Fig. 1.12. Equilibrium between linear and cruciform architectures at the site of a palindromic sequence in double-stranded DNA

transcription, RNA polymerase appears to generate on the DNA array a negative superhelicity upstream of the polymerase and a positive superhelicity downstream (see Fig. 1.11).

As a secondary effect, the unwinding of the DNA downstream of the polymerase induces a positive superhelicity, while the winding upstream produces a negative superhelicity. As mentioned in the last section, the negative superhelicity generated upstream of the polymerase favours the transition from a B conformation to a Z conformation in the presence of the right sequences, e.g., poly d(G-C), and this probably helps to minimise structural stresses in the transcription machinery. Note that a negative superhelicity favours transcriptional activity. Indeed, the RNA polymerase binds very well on circular DNA with a negative superhelicity but only weakly on relaxed circular DNA. This is one reason why transcription occurs near nucleosomes.

One of the last points to consider is the fact that stresses induced on the helix by a positive or negative superhelicity can generate major structural changes in the helix as a secondary effect. This is the case with regard to the formation of cruciform structures when there are palindromic sequences (see Fig. 1.12), structures revealed in supercoiled circular DNA [20].

For palindromic sequences, the linear form is in equilibrium with the cruciform structure. Note that this equilibrium transiently generates single-strand structures sensitive to SI nuclease. The shift in equilibrium depends on a certain number of parameters such as the presence of magnesium and the degree

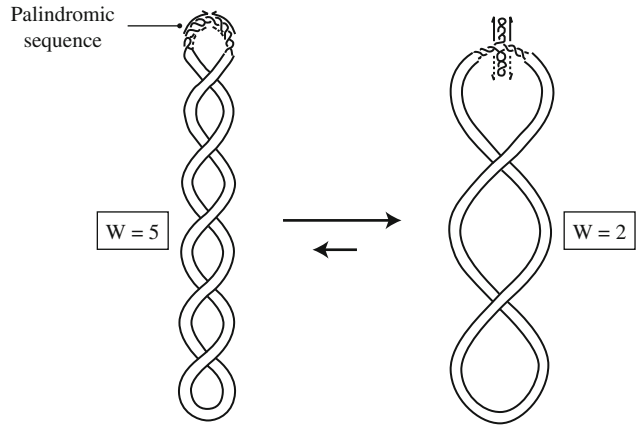

Fig. 1.13. Relation between the degree of supercoiling of DNA and generation of a cruciform structure from a palindromic sequence

of superhelicity. High superhelicity induces formation of the cruciform structure, whereas relaxation favours the linear form. In actual fact, formation of the cruciform structure induces as a consequence a reduction in superhelicity (see Fig. 1.13), and in the end reduces the free energy of the system. In the example of Fig. 1.13, if the twist T is considered to remain constant, then formation of the cruciform structure reduces the linking number L: $\Delta L = 3$.

It should be stressed that the role of these cruciform structures remains rather mysterious. Although it is easy to identify cruciform structures in vitro in circular DNA with a high level of superhelicity, it is extremely difficult to identify such structures in the genomes of eukaryotic cells under physiological conditions. However, the use of monoclonal antibodies directed against these structures and also photoinduced crosslinking experiments suggest that cruciform structures do indeed exist in living cells [21].

For other reasons, their presence as transient structures would appear highly probable:

- Type I and II topoisomerases, enzymes involved in the relaxation of superhelicity, preferentially recognise cruciform structures and cut DNA in the vicinity of these structures. In the light of this observation, cruciform structures can be considered as topological stress markers.
- Palindromic sequences are often found in sequences regulating gene expression (promoter sequences), the cruciform structures generated by these palindromes being specifically recognised by proteins regulating gene expression (transcriptional activators or repressors). This last point suggests that cruciform structures in relation with topological DNA variants might play an important part in regulating genetic expression.

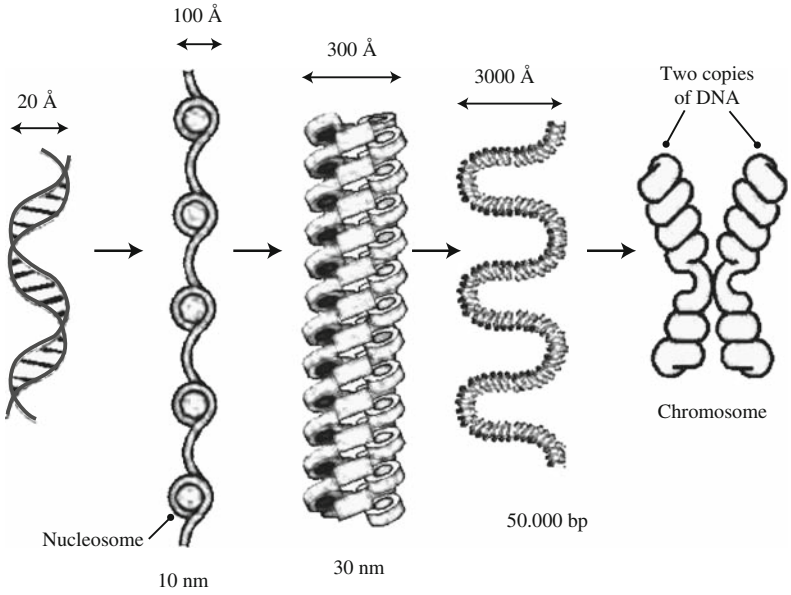

Fig. 1.14. Organisation and packing of DNA in chromatin and chromosome

1.4 Methylation of DNA

In association with geometric and topological modification of DNA, another essential ingredient in the control of genetic expression is chemical modification of the bases themselves, and in particular the methylation of cytosines.

Before discussing this point in more detail, it will be useful to review briefly the way DNA is organised in the nuclei of eukaryotic cells. The architectural organisation of DNA in the nucleus, and the eventual formation of chromosomes, are largely dictated by the need for a closely packed structure. Indeed, the human genome comprises some 3 billion base pairs in the form of double helices. In linear form, this would have a length of about two meters. The problem here is that these two meters of helix must somehow be packed into the nucleus, which measures only about one micron in diameter. Figure 1.4 summarises the organisation of the DNA and the different levels of packing.

The first level of organisation is the association of DNA fragments, containing some 200 base pairs each, with a globular protein structure comprising four histone dimers H2A, H2B, H3 and H4 (a histone octamer) to form the nucleosome, a structure mentioned in the last section (see Figs. 1.8 and 1.14). Note that the wrapping of DNA around the histone octamer is stabilised by H1 histone, which fixes the structure into place (see Fig. 1.8). The chain of nucleosomes thereby formed constitutes a 'fibre' of diameter about 10 nm (see Fig. 1.14). Under biological conditions, this fibre folds up to form the chromatin fibre of diameter 30 nm. The latter is then compacted in the form of

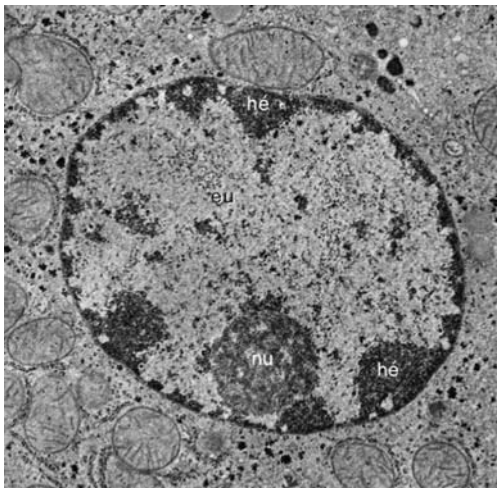

Fig. 1.15. Electron microscope image of the nucleus of an epithelial cell showing the non-uniformity of the chromatin. He: heterochromatin regions where DNA is more densely packed and hence more opaque to electrons. Eu: euchromatin regions where the DNA is less densely packed. Nu: nucleolus

a chromosome before cell division. To a large extent, DNA compaction is effected by electrostatic forces resulting from interactions between negatively charged phosphates on the DNA and basic amino acids on the positively charged histones.

In fact, within the nucleus of a non-multiplying cell, the degree of close-packing of the DNA is variable from one point to another. There are regions where it is less densely packed called euchromatin and others where it is more densely packed called heterochromatin (see Fig. 1.15). It was shown very early on that the heterochromatin with its densely packed DNA and low genetic expression is characterised among other things by a high level of methylation of the DNA, whereas the euchromatin, a region of strong genetic expression, is usually not significantly methylated and not very densely packed. Despite this kind of observation, DNA methylation was originally viewed by the scientific community as a secondary phenomenon of little biological interest.

However, major progress over the last few years has shown that methylation does in fact play an essential role in several biological processes associated with development, such as deactivation of the X chromosome in female mammals, genomic imprinting, and the expression of genes specific to different tissues. It is now accepted that the methylation of cytosines, in the context of the architectural organisation of chromatin, is one of the key mechanisms for regulating gene expression. This chemical modification of DNA is said to be epigenetic, because it can be transmitted from one cell to another and modulates the activity of a gene without directly affecting the sequence. The methylation of cytosines is preferentially located in specific genomic regions

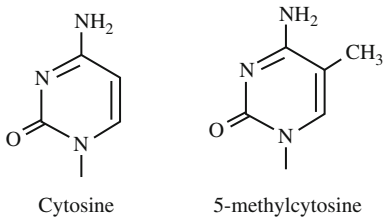

Fig. 1.16. Chemical structure of cytosine and 5-methylcytosine

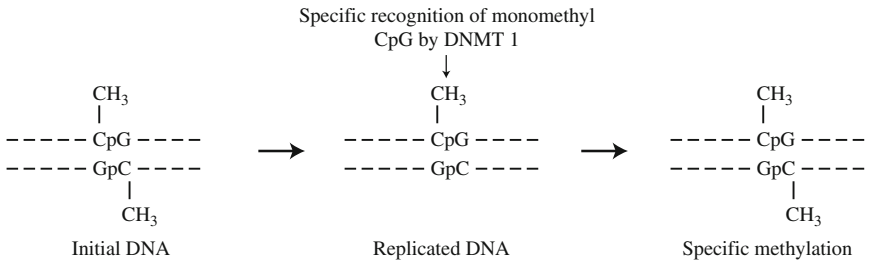

Fig. 1.17. Conservation of methylation in CpG sequences during DNA replication

within CpG dinucleotides. Indeed, globally speaking, 2–7% of cytosines in DNA are methylated, while some 70–90% are methylated in CpG sequences.

Finally, it seems more and more obvious that an exaggerated level of methylation in DNA plays a fundamental role in cancerogenesis. Such hypermethylation could deactivate genes suppressing tumour formation, thereby leading to tumorigenesis comparable to that induced by genetic mutation [22].

1.4.1 Methylation of Cytosine

In eukaryotes, the methylation of cytosine at position 5 is catalysed by a methyl transferase (see Fig. 1.16). The catalysed reaction involves transfer of a methyl group from S-adenosyl-methionine to a cytosine built into the DNA double helix. There are in fact three families of methyl transferases: DNMT 1, 2, together with 3a and 3b. DNMT 1 seems to be more specialised in maintaining overall levels of DNA methylation during cell multiplication. DNMT 3a is the methylation enzyme for gene regulating sequences and DNMT 3b is more specifically involved in methylation of centromeric sequences. These two enzymes can catalyse methylation of CpG sequences de novo, whereas DNMT 1 catalyses the methylation of semi-methylated CpG sequences.

An important point here is that the double methylation observed in CpG sequences is a structural characteristic that can be transmitted during cell multiplication. This is related to the semi-conservative nature of DNA replication and the specific recognition of monomethylated CpG sequences by methyl transferases, in particular DNMT 1 (see Fig. 1.17).

The enzyme DNMT 1 involved in maintaining the methylation of cytosines in CpG sequences is a very large protein, containing 1,618 amino acids. This protein is characterised by the presence near its N-terminal of three recognition regions denoted RTR1, RTR2, and RTR3, where RTR stands for replication target region, which are DNA replication regions. The C-terminal region contains the catalytic region and the recognition region of the semi-methylated DNA.

It is worth stressing that the level of expression of DNMT 1 varies during the cell cycle with a maximum in the G1S and S phases (S = DNA synthesis). This corresponds to the need to methylate cytosines on newly synthesised DNA strands. Consequently, and rather unexpectedly, one may consider that a high potential for methylation is not only compatible with, but even favours active cell proliferation. In confirmation of this idea, high levels of DNMT 1 expression have been observed in acute and chronic myelogenous leukaemia [23]. Moreover it would seem that many tumour cell lines can be characterised by hypermethylation of tumour-suppressing genes, and this favours the growth of these tumours. Such hypermethylation is mainly observed in CpG islands (see the next section) preferentially located in promoter regions. The kind of methylation related to the tumour phenotype can in fact be characterised by an overall hypermethylation of the genome, associated with hypermethylation of CpG islands controlling the expression of tumour-suppressing genes and, more generally, genes involved in the negative regulation of cell proliferation.

It should be stressed that the epigenetic mechanism of gene regulation and expression is a rather complex system, involving several concomitant processes. As an example, the methylation of gene regulating sequences (promoters), in a context of inhibition of expression, is accompanied by methylation of histones, in particular H3, on lysine 9, this leading to a form of cooperation between two methylation processes. We shall see in Sect. 1.4.4 that bimethylated CpG sequences are specifically recognised by proteins such as MeCP2, which binds onto these sequences. The complex formed in this way subsequently recruits histone deacetylases (HDAC), which catalyse the elimination of acetyl groups present on the histones. Acetylation of the histones facilitates the action of chromatin remodelling factors, leading to architectural changes allowing the opening and activation of promoters. The regulation machinery is then completed by the binding of other regulatory proteins such as HP1 (heterochromatin protein 1). This machinery can be summarised as follows:

- **Active Region for Gene Expression:**
  - Euchromatin: lightly packed chromatin.
  - Little methylation of CpG sequences.
  - Little methylation of H3.
  - Absence of HP1 proteins.
  - High level of acetylation of histones (action of histone acetylases HAT).
- **Inactive Region for Gene Expression:**
  - Heterochromatin: densely packed chromatin.

- High level of methylation of CpG sequences.
- High level of methylation of H3.
- Presence of HP1 proteins.
- Little acetylation of histones (action of histone deacetylases HDAC).

1.4.2 CpG Sequences

As we have seen, cytosines in the CpG dinucleotide sequence are favoured targets for methylation. Interestingly, the CpG sequence is much less well represented in vertebrate genomes [24]. This statistical anomaly resulting from selection pressure is probably due to the fact that cytosine is very easily methylated to give 5-methylcytosine (see Fig. 1.16), which is then easily deaminated to give thymine. This leads to a guanine–thymine mismatch that is not recognised by repair systems. This under-representation of CpG sequences is also observed in mitochondrial DNA, exemplifying the adaptation of prokaryotic DNA to a eukaryotic environment and functionality.

Despite their comparative overall scarcity in the genomes of higher eukaryotes, CpG sequences play a major part in the control of gene expression, with methyl-CpG sequences displaying a significant repressive potential. One remarkable feature relating to the regulatory function of CpG sequences is the presence of high CpG concentrations (CpG islands) [26] in the vicinity of promoter sequences for genes essential to the functioning of the cell and usually constantly expressed. These islands, comprising more than 200 nucleotides, are characterised by a high density of GC bases and a low level of methylation of the cytosines in the CpG sequences (in contrast to what is observed in isolated CpG sequences). Such hypomethylation is a prerequisite for strong gene expression, since methylation usually corresponds to inhibition of gene expression. Note that over-representation of CpG islands, associated with hypermethylation, can lead to anomalous under-expression of genes downstream of these islands. This happens, in particular, in the mental retardation syndrome associated with the fragile X site, characterised by deactivation of the FRM1 (fragile X mental retardation 1) gene. Deactivation of FRM1 can be imputed to an increase in the number of CGG triplets downstream of the promoter for this gene. In the population as a whole, the number of repeats is somewhere between 5 and 59 CGG, whereas patients affected by this syndrome have more than 200 repeats [27].

In order to understand the role played by CpG sequences in the functional regulation of the genome, it is essential to study the structural characteristics of these sequences. Note to begin with that the G-C pairing is stabilised by three hydrogen bonds, in contrast to the pairing of A-T nucleotides which only involves two hydrogen bonds (see Fig. 1.18). The first gas-phase measurements of the binding energy [28], later confirmed by other techniques, gave values of 21 kcal/mol for the G-C pairing, compared with 13 kcal/mol for the A-T pairing.

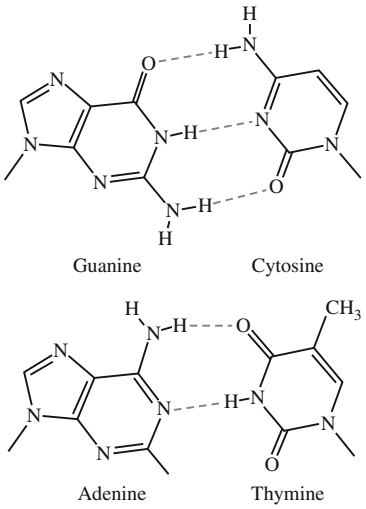

Fig. 1.18. Comparative pairing of G-C and A-T bases in a double-strand DNA molecule

The high binding energy of the G-C pairing makes the CpG dinucleotide sequence an extremely stable entity, which determines several of its characteristic properties. Apart from its stability, the CpG sequence diplays an extraordinary flexibility. As an example, in B helices it has been observed that elongation and unwinding are more energetically favourable in $d(CpG)_2$ (9.8 kcal) than in $d(GpC)_2$ (27.8 kcal). Note also the low twist angle (about 30°, compared with 40° for GpC sequences) characterising the geometry of stacks of base pairs in CpG sequences and the high positive roll which leads to an opening of the stacks of bases towards the minor groove of the helix. The geometric parameters of CpG sequences facilitate specific molecular recognition, in particular by methylases catalysing de novo methylation of CpG sequences.

1.4.3 Structure of Methylated CpG Dinucleotides

In the double helix, the two methyl groups symmetrically positioned on the cytosines of the CpG sequence are located in the major groove of the helix (see Fig. 1.19) and thus form a highly distinctive motif which can be specifically recognised, in particular by (effector) proteins carrying out genetic functions.

A lot of work has been done to investigate the effects of methylation on the local structure of the CpG sequence and adjacent sequences. To a first approximation, the main consequences of the presence of two methyl groups situated close to one another and protruding into the major groove are as follows:

- increased hydrophobicity of the groove,
- establishment of hydrogen bonds between protons of the methyl groups and amino acids of the protein ligands (see the next section),

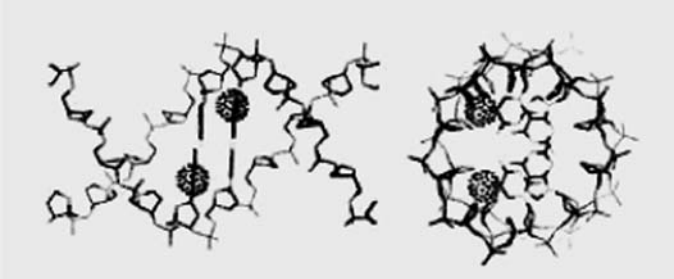

Fig. 1.19. Positioning of methyl groups in a CpG sequence symmetrically methylated on the cytosines. *Left*: View of methyl groups on the major groove side. *Right*: View in a plane perpendicular to the axis

- modified accessibility of the deep part of the groove, given the steric hindrance of the two methyl groups.

Regarding the DNA geometry, the general consensus is that methylation of a cytosine has little effect on the local structure of the helix. The main impact concerns the stability of the helix and molecular dynamics. Indeed, crystallographic studies [29] carried out on a d(ACCGCCGGCGCC) dodecamer have shown that the geometry of a duplex methylated on the central cytosine led to two new hydrogen bonds being set up between protons of the methyl group and oxygen atoms of the phosphates, with consequent stabilisation of the double helix. Moreover, NMR studies combined with molecular dynamics simulations [30] have shown that methylation of the CRE sequence (cAMP responsive element) $d(GAGATGAmCGTCATCTC)_2$ leads to the adoption of a BII conformation with reduced flexibility of the helix, notably in $5'$ adjacent sequences, inducing steric hindrance due to the methylated cytosine. These relatively modest structural and dynamic changes are enough to cause a drastic change in the interactions between the DNA and regulating proteins such as transcription factors.

1.4.4 Specific Recognition of Symmetric Methylation by Proteins

In vertebrates, there is a family of proteins that specifically recognise the symmetrically methylated CpG sequence. This family includes the protein MeCP2, already mentioned above, and also the proteins MBD1, MBD2, MBD3 and MBD4. These proteins share a methyl-CpG binding domain. This domain, located near the N-terminal region, comprises 70 amino acids [31].

From a biological standpoint, it is interesting to note that MeCP2 [32], after binding onto a methylated CpG sequence, is then able to recruit histone deacetylases. Deacetylation of the histone lysines releases positive charges and this favours electrostatic interactions between histone and DNA, thereby increasing the compaction of the chromatin and rendering it mute from a genetic point of view. When it binds to methylated CpG sequences, MeCP2

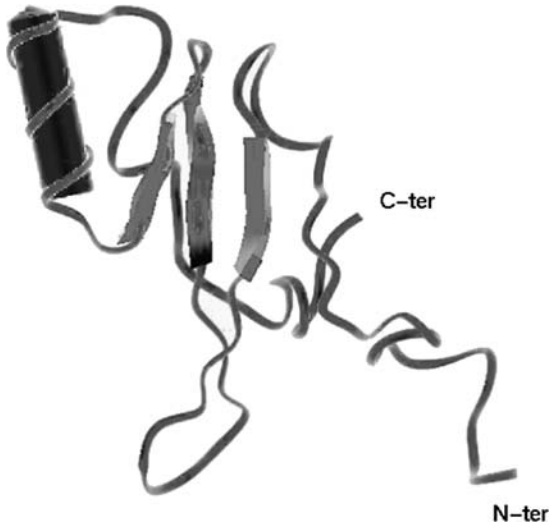

Fig. 1.20. Structure of the DNA binding domain of the protein MeCP2

acts as a key element in transcriptional repression. MeCP2 is assumed to bind onto a single methyl-CpG motif via essentially hydrophobic interactions. In fact the molecular mechanism for recognition of methyl-CpG sequences by the MeCP2 protein is not yet fully understood. However, the three-dimensional structure of the DNA binding domain is now known (see Fig. 1.20).

This domain is characterised in particular by a low level of structuring and a consequently high flexibility. This flexibility allows the protein to adapt itself to the rigidity of methylated CpG sites and to the steric hindrance caused by the presence of the two methyl groups. Mutagenesis and NMR studies [33] have shown that arginine-111, which interacts with aspartate-121, is one of the key amino acids controlling the specific binding of the protein to the methylated CpG sequence. Moreover, it would seem that the presence of sequences adjacent to the CpG site that are rich in AT base pairs favours a high-affinity binding of MeCP2 onto the CpG sequence.

In any case, through its binding on methyl-CpG sites, MeCP2 plays an essential role in the control of genetic expression. This is confirmed by the fact that mutations perturbing the binding of the protein onto methylated CpG sequences lead to the appearance of a pathology which mainly affects girls (Rett's syndrome), characterised by anomalous development of the central nervous system. This pathology, now considered to be a genetic disease affecting the X chromosome (the MeCP2 gene is carried by the X chromosome), is transmitted as a dominant character. The molecular etiology of this pathology clearly demonstrates the importance of epigenetic transcriptional repression processes.

1.5 Conclusion

The examples described in this chapter show the diversity and complexity of processes involved in regulating the expression of the genetic code. The most striking point is that DNA serves both as the physical support for genetic information and as the major regulator for reading this information. Changes in geometry, changes in topology, and changes in chemical structure all contribute in a concerted way, on the molecular scale, to the precise mechanisms regulating gene expression. Indeed, these modifications are signals triggering the mobilisation of effectors (mainly proteins) present in the environment of the DNA. These effectors are both regulatory elements (activators or repressors) and also synthesising elements carrying out the transformation of DNA code into RNA code (transcription), and then the transformation of RNA code into protein (translation).

Research over the last few years has revealed the fundamental role of epigenetic regulation of gene expression, i.e., regulation that is not directly linked to the gene sequence, and in this context, the equally fundamental role of repression processes, particularly those linked to DNA methylation and the compaction of chromatin structures (DNA plus proteins). Such processes repressing gene expression are probably the key features of development and cell differentiation. One emerging feature is that anomalies relating to these repression mechanisms may lie at the origin of, or at least large contribute to the occurrence of many major pathologies. This is the case for tumour transformation, which seems to result from an anomalous repression of genes whose function is to inhibit cell proliferation and maintain cell differentiation. Some of these genes have been clearly identified as tumour-suppressing genes.

From this point of view, it is striking to observe that the level of methylation of the genome increases steadily with the age of the individual. This observation may throw new light on the relationship between the incidence of cancer and aging. There can be no doubt that one of the great scientific challenges in biology will be to clarify the epigenetic mechanisms regulating gene expression, and beyond this, in the case of anomalies, to find ways to act and restore these regulation processes, by pharmacological means if need be.

References

1. Watson, J.D., Crick, F.H.C.: Molecular structure of nucleic acids, Nature **171**, 737 (1953)
2. De Santis, P., Palleschi, A., Savino, M., and Scipioni, A.: Biochemistry **29**, 9269 (1990)
3. Bolshoy, A., McNamara, P., Harrington, R.E., Trifonov E.N.: Proc. Natl. Acad. Sci. USA **88**, 2312 (1991)
4. `http://rumour.biology.gatech.edu`
5. Djuranovic, D., Hartmann, B.: Conformational characteristics and correlations in crystal structure of nucleic acid oligonucleotides: Evidence of sub-states, J. Biomol. Struct. Dyn. **20** (6), 1 (2003)

6. Mirau, P.A., Kearns, D.R.: Unusual proton exchange properties of Z-form poly[d(G-C)], Proc. Natl. Acad. Sci. USA **82**, 1594 (1985)
7. Misra, V.K., Honig, B.: The electrostatic contribution to the B to Z transition of DNA, Biochemistry **35**, 1115 (1996)
8. Kawaga, T.F., Howell, M.L., Tseng, K., Ho, P.S.: Effects of base substituents on the hydration of B- and Z-DNA: Correlations to the B- to Z-DNA transition, Nucleic Acids Research **21**, 255978 (1993)
9. Peck, L.J., Wang, J.C.: Energetics of B-to-Z transition in DNA, Proc. Natl. Acad. Sci. USA **80**, 6206 (1983)
10. Le Ber, P., Schwaller, M.A., Auclair, C.: Effect of intercalative binding compared to external binding on Z/B equilibrium of poly-d(Gme5C) using fluorescent oxazolopyridocarbazoles as probes, J. Mol. Recognit. **2** (4), 152–157 (1989)
11. Ha, S.C., Lowenhaupt, K., Rich, A., Kim, Y.G., Kim, K.K.: Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases, Nature **437**, 1183 (2005)
12. Schwartz, T., Rould, M.A., Lowenhaupt, K., Herbert, A., Rich, A.: Crystal structure of the Za domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA, Science **284**, 1841–1845 (1999)
13. Osipiuk, J., Skarina, T., Edwards, A., Savchenko, A., Joachimiak, A.: 1.55 Å crystal structure of putative Z-DNA binding protein AF2008 from *Archaeoglobus fulgidus*, ACA05 W0243
14. Witting, B., Wolfl, S., Dorbic, T., Vahrson, W., Rich, A.: Transcription of human *C-MYC* in permeabilized nuclei is associated with formation of Z-DNA in three discrete regions of the gene, EMBO J. **11**, 4653 (1992)
15. Kwon, J.A., Rich, A.: Biological function of the vaccinia virus Z-DNA-binding protein E3L: Gene transactivation and antiapoptotic activity in HeLa cells, Proc. Natl. Acad. Sci. USA. **102**, 12759 (2005)
16. Champ, P.C., Maurice, S., Vargason, J.M., Camp, T., Ho, P.S.: Distributions of Z-DNA and nuclear factor I in human chromosome 22: A model for coupled transcriptional regulation, Nucleic Acids Research **32**, 6501 (2004)
17. Hamiche, A., Carot, V., Alilat, M., De Lucia, F., O'Donohue, M.F., Revet, B., Prunell, A.: Interaction of the histone (H3-H4)2 tetramer of the nucleosome with positively supercoiled DNA minicircles. Potential flipping of the protein from a left- to a right-handed superhelical form, Proc. Natl. Acad. Sci. USA **93**, 7588 (1996)
18. Wang, J.C., Jacobsen, J.H., Saucier, J.-M.: Physiochemical studies on interactions between DNA and RNA polymerase. Unwinding of the DNA helix by *Escherichia coli* RNA polymerase, Nucleic Acids Res. **4**, 1225 (1977)
19. Liu, L.F., Wang, J.C.: Supercoiling of the DNA template during transcription, Proc. Natl. Acad. Sci. USA **84**, 7024–7027 (1987)
20. Lilley, D.M.: The inverted repeat as a recognizable structural feature in supercoiled DNA molecules, Proc. Natl. Acad. Sci. USA **77**, 6468–6472 (1980)
21. Ward, G.K., McKenzie, R., Zannis-Hadjopoulos, M., Price, G.B.: The dynamic distribution and quantification of DNA cruciforms in eukaryotic nuclei, Exper. Cell Res. **188**, 235 (1990)
22. Baylin S.B., Herman, J.G.: DNA hypermethylation in tumorigenesis: Epigenetics joins genetics, Trends Genet. **16**, 168 (2000)

23. Mizuno, S., Chijiwa, T., Okamura, T., Akashi, K., Fukumaki, Y., Niho, Y., Sasaki, H.: Expression of DNA methyltransferases *DNMT1*, *3A*, and *3B* in normal hematopoiesis and in acute and chronic myelogenous leukaemia, Blood **97**, 1172 (2001)
24. Bird, A.P.: CpG-rich islands and the function of DNA methylation, Nature **321** (6067), 209 (1986)
25. Pollack, Y., Kasir, J., Shemer, R., Metzger, S., Szyf, M.: Methylation pattern of mouse mitochondrial DNA, Nucleic Acids Res. **12** (12), 4811 (1984)
26. Gardiner-Garden, M., Frommer, M.: CpG islands in vertebrate genomes, J. Mol. Biol. **196** (2), 261 (1987)
27. Fu, Y.H., Kuhl, D.P., Pizzuti, A., Pieretti, M., Sutcliffe, J.S., Richards, S., et al.: Variation of the CGG repeat at the fragile X site results in genetic instability: Resolution of the Sherman paradox, Cell **67**, 1047 (1991)
28. Sukhodub, L.F., Yanson, I.K.: Mass-spectrometric studies of binding energies for nitrogen bases of nucleic acids in vacuo, Nature **264** (5583), 245 (1976)
29. Mayer-Jung, C., Moras, D., Timsit, Y.: Effect of cytosine methylation on DNA–DNA recognition at CpG steps, J. Mol. Biol. **270** (3), 328 (1997)
30. Derreumaux, S., Chaoui, M., Tevanian, G., Fermandjian, S.: Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element, Nucleic Acids Research **29** (11), 2314 (2001)
31. Nan, X., Meehan, R.R., Bird, A.: Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2, Nucleic Acids Res. **21**, 4886 (1993)
32. Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F., Bird, A.: Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA, Cell **69**, 905 (1992)
33. Free, A., Wakefield, R.I., Smith, B.O., Dryden, D.T., Barlow, P.N., Bird, A.P.: DNA recognition by the methyl-CpG binding domain of MeCP2, J. Biol. Chem. **276**, 3353 (2001)

http://www.springer.com/978-3-540-88632-7

Nanoscience
Nanobiotechnology and Nanobiology
Boisseau, P.; Lahmani, M. (Eds.)
2009, XXXVIII, 1202 p. 530 illus., 30 illus. in color.,
Hardcover
ISBN: 978-3-540-88632-7