HISTORICAL OVERVIEW

The elucidation of the structure, role and how the nucleic acid molecules exert their function is an example of the fruitful combination of genetics, biology and chemistry. The knowledge of the structure and chemistry of the nucleic acids forms a crucial basis for the molecular biology and bio-engineering.

1. A particulate factor of inheritance.

Johann Mendel, son of a farmer, became a monk and adopted the name of Gregor. He was send to the University of Vienna to become a teacher, but he failed. His examinator described him as "*someone lacking insights and the ability to formulate his thinking and idea's*". When he returned to the monastery of Brno in 1853 at the age of 31, he was a failure. He belonged to the 'Augustinians' a monastic order which spread to the villages for teaching and education of the people. As he was not allowed to teach, he returned to the things he knew, and started growing plants and vegetables. He decided on his own to perform practical experiments in biology. A courageous thing to do as the bishop forbids the teaching of biology at that time.

At the time he started his experiments, there was great confusion about heredity. The hold idea was that the progeny was a mixture of both parents (remember half-blood/full blood). Probably when experiments were performed in which a recessive characteristic appeared, it was discarded and treated as an artefact. Mendel started his experiments in 1856, and spend some 8 years to complete the study. The plant which was carefully chosen was the pea. He took 7 characteristics to compare and follow:

- the shape of the seeds,
- the colour of the seeds,
- colour of the pods,
- shape of the pods,
- colour of the flowers,
- position of the flowers on the stalks,
- the length of the plan

By breeding and crossbreeding and counting the numbers of each progeny type resulting from each mating he formulated his first and second law. He even had an explanation for his observation. He suspected that the special characteristic of a particular species is determined by two 'particulate factors' (called genes nowadays). Every parent gives one of these two 'particulate factors' to his progeny. If the two factors from both parents are mixed, then one is dominant and the other is recessive.

Mendel published his results in 1866 in 'Journal of the Brno Natural History Society', and although this journal was available in many university libraries, nobody took interest in it. He even send a copy to prof. N. Geli and to other eminent scientists, but his experiments and results remained unknown.

The experiments of Mendel are astonishing. Nobody could have devised such experiments without having a firm idea about the outcome. To start with, Mendel chose 7 traits which appear to be located on different chromosomes. Amazingly, a pea has only 7 chromosomes (remember that the notion of chromosomes was unknown at that time). Secondly, as mentioned before, he worked at a time where it was common belief that crossing always had to result in a mixture of both parents. So, Mendel should have noticed that this was wrong. Possibly he realized that the sex of animals is an all or none heredity.

The bishop and the other monks were not very enthusiastic or pleased about the work of Mendel. The fact that Mendel admired the work of Darwin made it even worse. The first thing done, when Mendel died in 1884 the monks burned all his papers. The name and work of Mendel was completely forgotten until 1900.

2. Discovery of the 'nucleic acid' substance.

Friedrich Miescher (1844-1895) discovered in Tübingen a substance purified from pus cells which he obtained from the surgical bandages of nearby hospitals. After digestion with pepsin and treatment with hydrochloric acid and shaking with ether he prepared a crude extract of nuclei. From these nuclei he isolated an unknown compound which he called "nuclein" which is acidic in nature, soluble in dilute alkali and insoluble in dilute acid. Furthermore the compound contained high amounts of phosphorous. Here again, the work was done in the laboratory of prof. Hoppe-Seyler, who refused to publish the work until two students were able to repeat the experiments some two years after the discovery. The work was published in 1868,

two years after Mendel's publication. Later, when Miescher returned to Basel, Switzerland, it was shown that such substance was also present in the sperm head of the 'Rhine Salmon'. This nuclein was of high molecular weight and the phosphorous content totals 9.59 %.

Altman (1889) was the first to use the term 'nucleic acid' for his preparation of protein-free nucleic acids isolated from animal tissues and from yeast.

Neuman (1899) further improved and developed the method to prepare a protein-free sample of nucleic acids.

3. Further recognition of the 'nucleic acid' constituents.

Piccard (1874) discovered the purine bases in nucleic acids. From salmon sperm cells he extracted and isolated guanine and hypoxanthine with boiling hydrochloric acid.

Kossel's work (1879 until 1888) lead to the isolation of adenine and xanthine. Thymine was already isolated by Miescher but this pyrimidine was not properly identified until the work of Kossel and Neuman in 1894. Likewise cytosine was isolated and identified in 1902-1903 by Kossel, Studel & Levene. Ascoli isolated in 1900 uracil from yeast nucleic acid.

4. Rediscovery of the work of G. Mendel.

The classical laws of inheritance and segregation resulting from the pioneering work of G. Mendel and published in 1866, was rediscovered independently by three scientists, H. de Vries, C. Correns and E. Tschermak, around 1900 (16 years after Mendel's death). The essence of Mendels discovery was that hereditary traits or 'particulate factors' are independent of one another and each is transmitted as a separate unit from a parent to the offspring.

5. Genes reside on chromosomes.

It was found that a discrete number of threadlike particles, chromosomes, appeared during the process of cell division. Walter Sutton (1903) realized and proposed that the chromosomes corresponded exactly to the properties ascribed to Mendel's particulate units of inheritance.

The proof that genes lie on chromosomes required the demonstration that a particular gene is always present on a particular chromosome. This proof was provided by Morgan in 1910, when he described his experiments with the fruit fly *Drosophila melanogaster*.

Morgan proposed that the cause of genetic linkage is the *'simple mechanical result of the location of the factors in the chromosomes*'. This coincided with the idea that the genetic material needs to be on a long or large molecule with high molecular weight in order to contain all the required hereditable information. Note that the genes can be linked. The acceptance of the chromosomes harbouring the genes didn't solve the question about the nature of the genes, were they composed of proteins or DNA? This remained a matter of dispute.

6. DNA/RNA confusion.

Although nucleic acids were originally thought to be essential nuclear constituents, the occurrence of pentose type in the cytoplasm was suspected as long ago as 1905. However by 1930, a definite picture had emerged of the existence of two different types of nucleic acids:

- Nucleic acid of yeast yielded on hydrolysis A, G, C, and U and phosphoric acid and a sugar recognised by Hammarsten, Levene & Jacobs as ribose.
- Nucleic acids of thymus yielded A, G, C, T and phosphoric acid and a sugar at first thought to be hexose, but later shown by Levene to be deoxypentose = D2 deoxyribose. The following type, occurrence and name of the different nucleic acids was put forward:

Type	Ribonucleic acid	Deoxyribonucleic acid
Occurrence	Plants	Animals
Name	Phytonucleic acids	Zoonucleic acids

but this statement was not free from objection.

Brachet (1940) demonstrated the presence of pentose nucleic acids in amphibians, and in the anterior pituitary of rat and guinea pig and in toads eggs by histochemical means. Caspersson concluded from his spectrophotometry experiments that a high concentration of pentose nucleic acid was characteristic of cells in which rapid protein synthesis was taking place.From the experiments of Brachet and Caspersson it was concluded that both types of nucleic acids are present in all types of cells (plant and animal) and the main biological distinction between pentose and deoxypentose nucleic acids is that the former is mainly cytoplasmic and the latter almost exclusively nuclear. To this end, the names chromonucleic acid and cytonucleic acids were proposed.

Davidson & Wymouth (1943) were able to demonstrate by chemical methods that the pentose nucleic acid was not peculiar to embryonic tissues as it was also found in a corresponding series of adult tissues. Although total nucleic acid was higher in embryo than in adult tissue cells, it is the amount of pentose nucleic acid relative to deoxypentose nucleic acid varied from tissue to tissue and the deoxypentose nucleic acid was of the same order in embryonic as in corresponding adult tissue.Pentose nucleic acid of sheep liver was present in 3 or 4 times the amount of deoxypentose nucleic acid. The ribose was found to be conclusively of the D-ribose type.

7. Genetic material is DNA.

Transformation of pneumococcus bacteria was obtained by mixing killed, virulent, smooth bacteria with live, rough, avirulent pneumococcal bacteria. Infection of a mouse with this mixture could kill the mouse. In this case, smooth virulent bacteria could be isolated from the dead mouse as reported by Griffith (1928). It means that some property of the dead bacteria can transform the live bacteria. Avery showed in 1944 that the transforming activity resides in the DNA. Shortly after the DNA involvement was confirmed by the inability to transform when the purified DNA was treated with an enzyme that degrades DNA. Surprisingly at the time these experiments took place it was not even known that pneumococcus contained DNA, Avery and his colleagues were talking about a substance of which the elementary analysis conforms very closely to the theoretical values of pure DNA.

'The inducing substance, on the basis of its chemical and physical properties, appears to be a highly polymerised and viscous form of DNA. On the other hand the type III capsular substance, the synthesis of which is evoked by this transforming agent, consists chiefly of a non-nitrogenous polysaccharide. Thus it is evident that the inducing substance and the substance produced in turn are chemically distinct and biologically specific in their action and that both are requisite in determining the type specificity of the cells of which they form part.'The experiments of Avery were however difficult to accept, it was argued that transformation might reflect some particular role of DNA in capsular polysaccharide formation. Hershey & Chase (1952) produced T2 viruses with ^{35}S labelled proteins and ^{32}P labelled DNA. After incubation of bacteria with these viruses, followed by removal of the un-adsorbed phages, it appears that the ^{32}P labelled

DNA enters the bacterial cell upon infection and produces consequently virus particles, while the ^{35}S labelled proteins remained outside the bacteria.It is interesting to note that these experiments were immediately accepted although they are not as precise as the transformation experiments of Griffith and Avery. The changed climate of opinion had much to do with the immediate acceptance of the Hershey-Chase experiment compared with the disbelief in Avery's work.

8. Structure of DNA.

Chargaff (1950) pointed to certain regularities in the base composition of nucleic acids. Many nucleic acids differed in composition as regards molar proportions of the bases according to the biological source of the material from which they were derived. Even differences exist within one cell if one looks at the nucleus - cytoplasm distribution.

The concentration of the 4 nucleotides varies from species to species and

[A] = [T]

[G] = [C]

After several proposals for the structure of DNA, Watson & Crick proposed a structural model of DNA in 1953. Up to now, this model turned out to be correct. In part, this proposal was based on X-ray studies of DNA fibres by Wilkins and Franklin which were consistent with a helical conformation, but in large measure it was suggested by observation of the frequency of occurrence of the purine and pyrimidine bases. The structure in itself explains many of its characteristics (stability, propagation, coding capacity, involvement in transcriptional activation of genes).

9. The flow of genetic information.

In 1961 Jacob and Monod proposed the concept of a short-lived mRNA (messenger RNA) molecule to carry the genetic information from the nucleus to the cytoplasm where it is translated with the help of rRNA (ribosomal RNA) and tRNA (transfer RNA) into proteins. It was further demonstrated that this mRNA associated with bacterial ribosomes and became the template for protein synthesis. Thus the concept of the flow of genetic information was symbolized and became known as the central dogma:

$$\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{PROTEIN}$$

STRUCTURE OF NUCLEIC ACIDS

CONSTITUENTS OF NUCLEIC ACIDS

The complete hydrolysis of high molecular weight nucleic acids yields pyrimidine and purine bases, a sugar component and phosphoric acid.

1. The bases: pyrimidine and purine.

The nitrogenous bases fall into two types: pyrimidines and purines. Pyrimidines have a six-member ring; purines have fused five-member imidazole and six-member pyrimidine rings. (Note that the numbering of the purine ring atoms differs from that used for the pyrimidine ring.)

Each nucleic acid is synthesized mainly from only four types of base. The same two purines, adenine and guanine, are present in both DNA and RNA. The two pyrimidines in DNA are cytosine and thymine; in RNA uracil is found instead of thymine. The only difference between uracil and thymine is the presence of a methyl substituent at C5. The bases are usually referred to by their initial letters; so DNA contains A, G, C, T while RNA contains A, G, C, U. Certain modification of the bases may occur after their incorporation into nucleic acids. The pyrimidine and purine bases can undergo keto-enol (C=O/COH) or amino-imino (NH2/NH) tautomerism. However, in the native nucleic acid at neutral or acidic pH it will be the keto and the amino form which are definitely predominant.

2. Sugars.

Two types of pentose, β-D-ribose and β-D-2-deoxyribose, are found in nucleic acids. These distinguish **r**ibo**n**ucleic **a**cid and **d**eoxyribo**n**ucleic **a**cid respectively, and give rise to the general names RNA or DNA for the two types of nucleic acid. The difference between the two pentoses lies in the absence/presence of the hydroxyl group at position 2' of the sugar ring.To avoid ambiguity between the numbering of the heterocyclic base and the sugar ring atoms, position on the pentose ring are given a prime ('). In nucleic acids, the ribose and deoxyriboses occur in the furanose form (furan is a five-membered oxygen containing ring). The orientation

of the OH group at the C1' in the ring is very important. Where the OH at the C1' and C4' are on opposite side of the ring are indicated as β. The D indicates the configuration of the centre of asymmetry of the most remote end from the aldehydic end of the sugar molecule.

3. Phosphoric acid

As we will see, the sugars are connected to each other in the polymerised RNA and DNA molecules. Two pentose rings are joined by a phosphoric acid molecule which is forming an ester bond with the 5' and 3' C. As a result at neutral pH (or at pH above 1) only one negative charge remains of the phosphoric acid. The pK value for a primary phosphate ionisation of a phosphor-diester has a value near 1.

4. Nucleosides.

When a purine or pyrimidine base is linked to ribose or deoxyribose the resulting compound is known as a nucleoside. Depending on whether A, G, C or U (T) are condensed with a ribose (deoxyribose) we talk about adenosine (deoxyadenosine), guanosine (deoxyguanosine), cytidine (deoxycytidine) or uridine (deoxythymidine). The ribonucleoside from hypoxanthine is named inosine.

In the pyrimidine nucleosides, the sugar and base are joined by a □-glycosidic link from the C1' of the pentose to the N1 of the pyrimidine base. In the purine nucleosides, the C1' of the sugar is connected through the β-glycosydic link to the N9 of the purine base. Digestion products from particularly tRNA yield small amounts of pseudo-uridine.

5. Nucleotides.

As mentioned before, the polymeric forms of RNA and DNA contain phosphoric acid esters of the nucleosides. These esters are called nucleotides. When DNA is broken into its constituent nucleotides, the cleavage may take place on either side of the phosphodiester bonds. So, two types of nucleotides can be released: the nucleoside-3'-monophosphate and the nucleoside-5'-monophosphate. Since the ribonucleosides have three free hydroxyl groups on the sugar ring, three possible ribonucleosides can be formed.

The ribonucleoside 5'-phosphates may be further phosphorylated at position 5' to yield 5'-di- and -tri-phosphates. Thus, adenosine 5'-phosphate (AMP) yields adenosine 5'-diphosphate (ADP) and adenosine 5'-triphosphate (ATP). The bonds between the first (α) and the second (β) and between the second (β) and the third (γ) phosphate groups are energy rich and are used to provide an energy source for various cellular activities. The triphosphates are the forms from

which the nucleic acids are synthesized. The deoxyribonucleoside 5'-phosphates are referred to as dAMP, dADP, dATP, etc... Cyclic 2',3' monophosphates can be formed on hydrolysis of RNA molecules, and ribonucleoside 3',5' cyclic monophosphates of adenine and guanine occur in many tissues where they play multiple roles in the regulation of metabolic pathways.

THE STRUCTURE OF DNA

1. The poly-deoxynucleotide chain.

Deoxynucleic acids are polynucleotides in which the phosphate groups link the C5' of one sugar to the C3' of the next. This means that a single stranded DNA chain has a polarity. By convention, we write the sequence of such a polymer starting with the nucleotide attached to the sugar with the free C5' end.

2. The double helix.

The discovery of the double helix structure of the DNA is an important milestone in the history of molecular biology. Two notions converged in the construction of this model proposed by Watson & Crick in 1953 (*Nature*).

First, X-ray diffraction data of extended DNA fibres showed that DNA has the form of a regular helix, making a complete turn every 34 Å, and with a diameter of about 20 Å (10 Å = 1 nm). Since the distance between two adjacent nucleotides is 3.4 Å, this means that there must be 10 nucleotides per turn. The density of the DNA was in agreement with an association of two individual polynucleotide strands.

The second critical feature that explained how this form of organization is accomplished lay in the previous observation of Chargaff that, irrespective of the actual amounts of each base, the proportion of G and C is always the same in DNA, and the proportion of A and T is always the same. Thus any DNA can be characterized by its (G+C)/(A+T) ratio, which ranges from 25 to 75 %, but is a characteristic of each species. Watson and Crick proposed that the two polynucleotide chains in the double helix are not connected by covalent bonds, but associate through hydrogen bonding between the bases. These hydrogen bonds can be formed because of the proper location of the proton donor and proton acceptor atoms. The structures of the thymine with adenine are complementary in that they can be fitted together in the same plane so that two hydrogen bonds can be formed between them. At the same time, the atoms by which the bases are attached to their sugar molecules, the N1 of thymine and the N9 of adenine, are at opposite ends of the molecular complex. The same situation holds for cytosine and guanine, except that the base association in this case is accompanied by the formation of three hydrogen bonds. A very significant point is that the "end-to-end" distances are nearly the same for the A·T and G·C pairs. The suggested base-pairing is in accord with a content of G = C and A = T, as established for a large number of DNA samples.

The base pairs are essentially flat and may be stacked one above the other like a pile of plates so that the molecule is readily represented as a spiral staircase with the base pairs forming the treads. The two polynucleotide strands are of opposite polarity in the sense that the terminal nucleotide of one chain has a free 5' end whereas the complementary chain has a free 3' end at that same terminal end of the molecule. This means that the two strands are running in opposite directions, and that by the sequence of one strand determines immediately the nucleotide sequence of the other, complementary strand. From the results of the nearest neighbour analysis it can be proved that the two strands must have opposite polarities.

Each base pair is rotated by some 36° around the axis of the helix, relative to the next base pair. Thus 10 base pairs make a complete turn of 360°. The twisting of the two polynucleotide chains around each other forms a double helix. The handedness of this helix is right handed, i.e. the turns run clockwise looking along the helix axis. The examination of the structure of the double helix of DNA shows that there are two grooves along its surface: these are known as the wide and narrow grooves. In B-DNA the wide groove is about 22 Å across and the narrow groove approximately 12 Å across. As we will see, the width and the accessibility of these grooves changes in deviating from the B-DNA.It should be noted that always the same side of the base pairs are exposed in the narrow or minor groove, while the other side of the base pair will always be present in the wide or major groove. The base-sugar linkage has a defined orientation, it is normally *anti* with the C8 of the purine or the C6 of the pyrimidine ring over the sugar. However, in some special DNA structures (Z-DNA) the *syn* position has been observed. The sugar ring is puckered, with either the C2' or the C3' being displaced from the planar conformation in the direction of the C5'. Usually the sugar is in the C2' *endo* configuration. (In the *exo* conformation, one C' will be displaced in the opposite direction of the C5'). It can be seen that the phosphate group linkages contain a great flexibility.

3. Why a helix?

DNA inside a cell is surrounded by water. Sugars and phosphates are highly soluble in water, however the bases, adenine and thymine for example, can be regarded as insoluble in water at neutral pH, they are hydrophobic.

Therefore the bases will try to escape from the water. In a polydeoxynucleotide chain, you will find a water soluble region at one longitudinal end of the molecule and a more hydrophobic region at the other end of the molecule. When two polynucleotide chains encounter each other they will tend to approach each other with the water repellent parts of their molecule, leaving the sugar phosphates at the outside.From the thickness of the bases, and the distance in space of two consecutive phosphates we can see very easily why two polynucleotide chains should skew the ladder or form a helix. Nature has chosen the second option to remove the spaces between the bases. We can now calculate roughly the expected number of base pairs per turn. The phosphate to phosphate angle should be 2 x arcs in (2.5/9.0) = 32.3°. Thus each phosphate to phosphate rotation makes an angle of 32.3°/360° = 1/11 part of a circle, and that is why we have 11 phosphates to represent a complete turn of DNA. In DNA it turns out that we always have between 10 and 12 phosphates per turn of helix within each strand.

4. Alternative DNA structures.

The existence of at least three different structural forms of DNA has been observed by X-ray for a long time, and transitions between them occur when appropriate changes are made in the environmental conditions. The B-DNA form for which Watson and Crick constructed their model is found in fibres of very high relative humidity (92 %) and in solutions of low ionic strength. Therefore it was assumed that this form prevails in the living cell.

The A-DNA form is found in fibres at 75 % relative humidity and requires the presence of sodium, potassium, or cesium as the counter-ion. Instead of lying flat, the bases are tilted with regard to the helical axis (20°). The consequences of this difference are profound. In B-DNA the local helical axis is in approximately the same direction as the global double helical axis, but in A-DNA there is a roll angle of +20° (see further for definition of roll) between each successive base pair, which opens the minor groove and means that the local helical axis follows a superhelical path around the global double helical axis. The different geometry is also reflected in the disposition of the base pairs relative to the DNA grooves. Whereas in B-DNA the base pairs are centrally placed in the double helix, in A-DNA they are displaced outwards towards the minor groove, which is consequently narrowed and shallower relative to B-DNA. In addition the rise for each double helical turn differs for these two forms, 34 Å for B-DNA and

27-28 Å for A-DNA; and there are more base pairs/turn in A-DNA. The sugars in A-DNA molecules have the C3' *endo* conformation.

The A-form is biological interesting because it is probably very close to the conformation adopted by DNA-RNA hybrids or by RNA-RNA double-stranded regions. The reason is that the presence of the 2' hydroxyl group prevents RNA from lying in the B-form. Besides organic solvents or salts, the DNA can be forced into the A form by proteins.

The C-DNA form occurs when DNA fibres are maintained in 66 % relative humidity in the presence of Li ions. It is also a right handed structure with fewer base pairs per turn than B-DNA. As in the A-DNA conformation the bases are not perpendicular to the helical axis, but are tilted by 6°.

The fine structural details at atomic resolution of the sugar phosphate chain could not be obtained from the X-ray diffraction studies on DNA fibres. More recently, with the advent of chemical DNA synthesis, oligonucleotides of unique sequence and defined length could be produced in large quantities and crystallized. These single crystal diffraction patterns confirmed the right-handed double helical structure and revealed unequivocally the position of each atom. From these analyses it is clear that the DNA is certainly not a smooth, regular double helix, but rather the helix shows important irregularities which are dependent on the underlying nucleotide sequence. The deformation affects not only the interaction with the surrounding water molecules but of course also with proteins.

Besides the deviations from a regular helix, a surprising structure was found for a single crystal of a particular oligonucleotide. Under condition of high salt concentration, DNA with alternating purine-pyrimidine sequence such as $(GC)_n$ apparently adopts a left-handed helical structure. It has the most base pairs per turn (12) in comparison to A, B and C- DNA, and so has the least twisted structure (-30°). Its overall structure is very skinny and the name 'Z-DNA' is taken from the zigzag path of the sugar-phosphate backbone along the helix. The negative charged phosphate groups lie close together in space, and need to be countered by the salt ions to decrease the electrostatic repulsions. The structural repeating unit is the dinucleotide. In case of the $(GC)_n$ Z-DNA the purine sugar linkage is in the *syn* conformation, the pyrimidine in the *anti*. Because of this, the normal base pair stacking cannot occur, but the pyrimidines can stack with pyrimidines on the opposite strand. The pucker of the sugar rings is changed so that, although

the sugar attached to the pyrimidine is now C2' *endo*, that to the purine is C3' *endo*. There is one deep groove between the sugar phosphate backbones corresponding to the minor groove in B-DNA.

As well as forming at high salt concentrations, Z-DNA will form in 1 mM $MgCl_2$, if the C5 of cytosine is substituted with methyl-, bromo- or iodo-groups. Oligo $(dG\text{-}d5\text{-}methylC)_n$ might occur in DNA inside the nucleus, this makes it seem plausible that Z-DNA could exist *in vivo* under the right circumstances. Z-DNA can also form in super coiled DNA plasmid molecules where the region of the left handed DNA serves to relax the tension of the super coiled molecule (see further). Moreover, a stretch of $(GC)_{26\text{-}32}$ can convert to Z-DNA while the regions on either side remain in the B-conformation. This strengthens the idea that conformational transitions could occur in vivo at specific sites. Also, naturally occurring nuclear proteins which bind to the Z-DNA form have been isolated. However, these suggestive evidences still did not prove that Z-DNA occurs inside the cell. Initial evidence for the existence of Z-DNA in cells came from the use of specific antibodies to Z-DNA. These antibodies bind to certain chromosomes of *D.melanogaster*, the polytene chromosomes. The reaction can be visualized because the chromosomes have an unusual structure, in which compact regions (bands) alternate with less compact regions (interbands). The Z-DNA resides in the interbands. It is not certain, however, to what extent Z-DNA forms during the required fixation processes which involve removal of basic proteins and which may induce super coiling and hence stabilize regions of Z-DNA. More recently, it has been shown that potential Z forming sequences interfere with the methylation and restriction of adjacent sequences and this is taken as showing that Z-DNA exists *in vivo*.

5. Special conformations

In solution, a DNA molecule does not assume a single static structure. Instead there is a dynamic flux of changes in axial and helical parameters. This means that in general the DNA molecule will adopt any preferred architecture in the absence of specific constraints. However, some particular sequences can adopt special preferred conformations which are sufficiently stable and rigid to resist transient structural changes.

DNA cruciforms

We discussed the lower stability of the TATA-like sequence due to the combined effect of its lower stacking interactions and lower amount of hydrogen bonds. These sequences readily unwind. Of course an unwinding would expose the bases with the surrounding water molecules. This can be prevented by reforming a double helix within one strand to form a *'cruciform'* configuration. It appears that the unwinding of DNA into a cruciform at TATA-like sequences is catalysed by partial unwinding of the DNA from 10.5 to about 12 base pairs per turn, or from T = 34° to about T = 30°. It has to be said that other palindromic sequences can also form a cruciform structure under superhelical stress, especially when they are interrupted by a few nucleotides which unwind easily. The same relaxation of superhelical DNA is accomplished by conversion of a short stretch from B DNA into Z DNA or single stranded DNA.

Intrinsic Bended DNA

The homopolymeric (dA)•(dT) tracts also belong to this class of sequences adopting a special static structure. It has long been apparent that these sequences are structurally distinct from random sequence DNA, adopting a helical periodicity in solution of 10 base pairs per turn in contrast to the average of 10.5-10.6 base pairs per turn. Furthermore, it is also exceptional in its failure to undergo the B to A transition in fibres, which is a property indicative of conformational rigidity. The crystal structure of a DNA dodecamer containing a run of six (dA)•(dT), was shown to be essentially straight, that is, the average planes through the base pairs themselves are parallel to each other and perpendicular to the helix axis. They have zero roll. On the other hand, the propeller twist is high, about 20 to 25°. This results in maximal overlap of the bases on each strand with a consequent increase in stacking energy and also the formation of a run of additional, non-Watson-Crick, cross-strand hydrogen bonds. These hydrogen bonds are located in the major groove and bridge the N6 position of adenine with the O4 of thymine. These positions thus have the potential for forming hydrogen bonds in two directions; the bonds are bifurcated. Both the increased stacking energy and the bifurcated hydrogen bonds would be expected to confer a conformational rigidity over and above that expected from base pairs with two hydrogen bonds.

The fact that the stretches of homopolymeric (dA)•(dT) runs are structurally distinct from DNA of random sequence makes the boundary peculiar. In the crystal structure, of d(CGCA_6GCG) the most abrupt changes in the direction of the local helical axis occur at the CpA step and at the GpC step 3' to the dA tract. At both these steps the normal purine clash is exaggerated by the difference in propeller twist, resulting in a large positive roll angle for the pyrimidine-purine step (CpA) and a corresponding negative roll angle for the purine-pyrimidine step (GpC). Therefore the helical axis bends as a result from these large roll angles at the junctions of the A•T tract with flanking base pairs. Since the angles at the 5' and the 3' ends of the tract are opposite in sign they would be additive when placed half a helical turn apart.It can be easily seen that a DNA molecule in which repeated stretches of (dA_5)•(dT_5) occur with their centres spaced at intervals of 10 base pairs, will lead to a coherent bending of this DNA. This intrinsic bended DNA was observed in kinetoplast DNA. Intrinsic bended DNA has a characteristic lower mobility in electrophoresis in polyacrylamide gels.

Besides the presence of intrinsic bended DNA in kinetoplast DNA where it probably aids in relieving the stress of the small DNA circles, bended DNA was also found in promotor regions. Moreover it has been shown that, in some instances, the presence of intrinsic bended DNA in itself could substitute for the effect of an activator protein bound to its target sequence.

Triple stranded helix

Until recently it would have been enough to learn just about Watson-Crick base pairs. But now it is important to learn about Hoogsteen base pairs as well, because such pairs show up occasionally in DNA. Karl Hoogsteen tried to confirm the Watson-Crick base pair for adenine and thymine by heating up a solution of these two bases and letting it cool slowly in order to make a crystal, but he found instead a different kind of base pair in his crystal. Apparently the adenine base has been rotated by 180° about the glycosidic bond in order to change between the two kinds of base pairing. The Hoogsteen and the Watson-Crick base pair for an A•T pair are roughly of equal stability: in both cases, there are two hydrogen bonds which hold together the two bases.

The Hoogsteen pairs for guanine and cytosine can also be drawn. There are two important differences however: first, the Watson-Crick G•C pair has three hydrogen bonds,

rather than two in the Hoogsteen conformation. So the Watson-Crick pairing for G•C is more stable than the Hoogsteen pairing in this case. Secondly, the Hoogsteen guanine-cytosine pair is only formed at low pH, since one of the nitrogens on cytosine must be protonated for this structure to form. The midpoint for protonation is pH 5, or slightly more acidic than the normal pH 7 to 8 found in cells. That is the main reason why practically all DNA double helices contain Watson-Crick rather than Hoogsteen pairs. The Hoogsteen G•C pair is not stable at neutral pH.

Sequences containing runs of (dCpT)•(dApG) or (dG)•(dC) are sometimes characterized by the acquisition of an asymmetric sensitivity to cleavage reagents which sense single DNA strands. This observation suggests that within this DNA conformation (called H-DNA) one strand, but not its complement, is freely accessible to such reagents. The proposed structure for this H-DNA is thus a region of triple stranded DNA in which a strand of the repeating sequence is wrapped in the major groove of a duplex of the same sequence. In this configuration the third strand can form Hoogsteen base pairs with the strands in the duplex that are parallel to itself, thereby stabilizing the structure. The existence of such a triple strand automatically means that the complement to the wrapped strand will not be involved in base pairing. Triple stranded H-DNA is essentially an asymmetrical structure which implies that, provided the Hoogsteen base pairing requirements can be satisfied, two isomers should, in principle, exist. For the homopolymer (dG)•(dC) this is indeed the case, the differences in the stability of the two forms being dependent on the pH and the presence of divalent cations. The pH effect can be directly related to the protonation of cytosine residues which allows the formation of Hoogsteen base pairs in one isomer but not in the other.

Three-stranded structures in nucleic acids were discovered soon after the discovery of double helical DNA. Sequences capable of adopting these structures were later mapped to the promotor regions of several eukaryotic genes, cloned into plasmids and shown to form triplexes under conditions of superhelical stress or low pH. More recently, intramolecular triple helix formation between oligonucleotides and double stranded DNA has been exploited for sequence-specific recognition of DNA sequences. Applications ranging from site-specific cleavage of chromosomal DNA to inhibition of transcription from specific genes have been demonstrated. Although no role for triplex formation in vivo has yet been found, a triplex DNA-binding protein

has been characterized from human cells. This protein may facilitate DNA triplex formation in vivo. DNA triplex structures might be formed between a Y•R double helix and a single stranded R strand. It was previously thought that Y*R•Y DNA triplexes would not form *in vivo* because of a requirement for low pH. However, it has now been established that polyamines, which are ubiquitous in living cells, promote Y*R•Y DNA triplex formation at neutral pH.

Quadruple helical DNA

Tetraplexes could theoretically be formed by the association of the two arms of a cruciform structure. A tetraplex DNA will also be generated in guanine-rich sequences such as occur in some promotors and also in telomeres.

Telomeres, the nucleic acid-protein complexes found at the ends of linear eukaryotic chromosomes, are believed to be responsible for the ability of linear chromosomes to replicate without shortening at the 5' end, to protect chromosome termini from degradation, and to be involved in chromosome organization, and anchoring to the nuclear envelope. They usually contain many tandem repeats of guanine-rich sequences. These extend in the 3' direction beyond the 5' end of the complementary strand, resulting in single stranded overhangs in telomeres from species as divergent as ciliates, yeasts, plants and humans. The folding topology of these overhangs has attracted much recent interest. *In vivo*, the G-rich segments form four-folded structures (tetraplexes or quadruplexes). Each G residue coordinates with two others from different segments, resulting in stacked G-tetrads.

The repeated tracts of 3 to 8 guanine residues are separated by short tracts of A or T in most, but not all telomeres. The A/T rich linking sequence then forms a loop linking the strands. These structures are very stable once formed, but form only at high DNA concentrations, and their biological relevance was uncertain until the recent demonstration that the □-subunit of the *Oxytrichia* telomere-binding protein catalyses G-tetraplex formation *in vitro*.

The topology of G-tetraplexes appears to vary depending on conditions. The G-rich segments of the *Oxytrichia* telomere sequence can form either a tetraplex in which all strands are antiparallel and the loops crossing over between the strands are lateral or, in solution, an alternative tetraplex with both parallel and antiparallel strands and diagonal loops. In a human telomeric DNA sequence, $(T_2AG_3)_3$, both lateral and diagonal loops were observed. Four

grooves are formed, one wide, one narrow and two medium. The combination of the loops and grooves should provide unique surfaces on human telomeres for ligand and protein recognition. Indeed multiple proteins have now been characterised that interact specifically with the chromosome ends.

STRUCTURE OF RNA

As explained before, the building up of an RNA primary structure is similar to the DNA: it consists of a polynucleotide chain with 5'-3' sugar-phosphate links. But of course the sugar is a ribose, and the thymine base has been replaced by uracil.

Why is 2'-Deoxyribose the Sugar Moiety in DNA? Common perhydroxylated sugars, such as glucose and ribose, are formed in nature as products of the reductive condensation of carbon dioxide we call photosynthesis. The formation of deoxysugars requires additional biological reduction steps, so it is reasonable to speculate why DNA makes use of the less common 2'-deoxyribose, when ribose itself serves well for RNA. At least two problems associated with the extra hydroxyl group in ribose may be noted. First, the additional bulk and hydrogen bonding character of the 2'-OH interfere with a uniform double helix structure, preventing the efficient packing of such a molecule in the chromosome. Second, RNA undergoes spontaneous hydrolytic cleavage about one hundred times faster than DNA. This is believed due to intramolecular attack of the 2'-hydroxyl function on the neighboring phosphate diester, yielding a 2',3'-cyclic phosphate. If stability over the lifetime of an organism is an essential characteristic of a gene, then nature's selection of 2'-deoxyribose for DNA makes sense. The following diagram illustrates the intramolecular cleavage reaction in a strand of RNA.

Structural stability is not a serious challenge for RNA. The transcribed information carried by mRNA must be secure for only a few hours, as it is transported to a ribosome. Once in the ribosome it is surrounded by structural and enzymatic segments that immediately incorporate its codons for protein synthesis. The tRNA molecules that carry amino acids to the ribosome are similarly short lived, and are in fact continuously recycled by the cellular chemistry.

Why did nature choose for uracil instead of thymine in RNA?

. The carbon atoms that are part of these compounds may be categorized as follows. All of these compounds are apparently put together from a three-carbon malonate-like precursor and a single high oxidation state carbon species. Such biosynthetic intermediates are well established. Thymine is unique in having an additional carbon, the green methyl group. Biosynthesis of this compound must involve additional steps, thus adding constructional complexity to the DNA molecules in which it replaces uracil. The reason for the substitution of thymine for uracil in DNA may be associated with the repair mechanisms by which the cell corrects damage to its DNA. One source of error in the code is the slow hydrolysis of heterocyclic enamines, such as cytosine and guanine, to their corresponding lactams. This changes the structure of the base, and disrupts base pairing in a manner that can be identified and then repaired. However, the hydrolysis product from cytosine is uracil, and this mismatched species must somehow be distinguished from the uracil-like base that belongs in the DNA. The extra methyl group serves this role nicely.

Secondary structure of RNA

Generally the RNA chain exists as a single polynucleotide chain rather than a double helix of antiparallel strands. Indeed, the base composition of an RNA molecule does not follow the Chargaff rules with A = T and G = C. However, as the bases on themselves are hydrophobic in nature, they tend to avoid the contact with water. They will try to base pair inter- or intra-molecular. The double helical region which is thus formed contains antiparallel strands kept together by 'tilted' Watson-Crick base pairing with an A-form structure, since the 2'OH group of the sugar hinders B structure formation.

When a sequence of bases is followed by a complementary sequence in the same chain, the polynucleotide may fold back on itself to generate an antiparallel duplex structure. This is called a hairpin. It consists of a base-paired, double helical region, the stem, often with a loop of unpaired bases at one end. In order to form a perfect hairpin a palindromic sequence is required. Of course such perfect palindromes occur only very rare and therefore the helical regions are usually not very regular. They show frequently unusual base pairing such as G•U, and non-paired residues may *'loop out'* or form a *'bulge'* on the stem.

When two more distant sequences of an RNA chain are complementary, they may come into juxtaposition to base-pair to form a double-helical region. Essentially this creates a stem with a very long single stranded loop. Also two different RNA molecules may have regions that are complementary and that base pair in the appropriate environment, usually these regions comprise rather short sequences.

Tertiary structure

The possibility exists that a stem has some complementary sequences to other unpaired regions. If these two parts of the molecule are brought together in space, then they can also be involved in secondary structure formation. This frequently aligns short stem regions which can be stabilized by stacking interactions. The structure is called a *pseudoknot* as long as the stem regions involved are shorter than a full helical turn.

As with DNA, short, triple stranded regions can occur in RNA in which two of the chains run parallel with one another. Such unusual structures have been studied with model systems where the homopolymer chains poly(A) and poly(U) have been shown to form a triple-stranded structure in which antiparallel poly(A) and poly(U) strands are held together by conventional Watson-Crick base pairing, whereas a second poly(U) strand uses Hoogsteen base pairs to bind in parallel to the poly(A) strand.

The RNA pseudoknots play important roles in many biological processes. In the simian retrovirus type-1 (SRV-1) a pseudoknot together with a heptanucleotide slippery sequence are responsible for programmed ribosomal frameshifting, a translational mechanism used to control expression of the Gag-Pol polyprotein from overlapping *gag* and *pol* open reading frames.

Types of RNA

RNA has a variety of functions within the cell and for each function a specific type of RNA is required. The types of RNA differ in chain length and in the secondary and tertiary structures.

RNA genomes

Some viruses contain a genome which is composed of RNA instead of DNA. The virus particles themselves contain sometimes duplex RNA (Reovirus). Other viruses such as the MS2 or polio virus pack the mRNA-like strand or the + strand, while others like the measles, flu and rabies viruses have a - strand (= strand complementary to the mRNA) RNA genome inside their particles.

Messenger RNA

From the physical separation between the genetic material in the nucleus and the occurrence of protein synthesis in the cytoplasm of eukaryotic cells, it was clear that DNA could not itself provide the template. It is the messenger RNA or mRNA that fulfils the function of intermediate molecule to transfer the genetic information to the ribosomes, the machinery of the protein synthesis.

Bacterial mRNA is unstable, and has only a brief existence. It is therefore difficult or even impossible to obtain the bacterial mRNA in an intact form and to use it in subsequent analyses.The bacterial mRNA contains two types of regions. The coding region consists of a series of codons representing the amino acid sequence of the protein, starting (usually) with an AUG and ending with a termination codon. But the mRNA is always longer than the coding region. Extra regions may be present at both ends of the coding region. Additional sequences at the 5' end that precede the coding region are described as leaders. Additional sequences that follow the termination signal and form the 3' end are known as trailers. Although part of the transcription unit, these sequences are not used to code for protein. The intercistronic regions that lie between the various coding regions of a polycistronic mRNA vary greatly in size. Actually the AUG startcodon may overlap with the UGA stopcodon of the previous coding region.

In eukaryotes, the search for the messenger also at first encountered difficulties. As with bacteria, mRNA constitutes only a small proportion of the total cellular RNA (roughly 3 to 5 % of the RNA mass). The first specific mRNA to be isolated was a globin mRNA from red blood cells. The mRNA's in eukaryotes are generally stable for a period of hours. They can therefore be isolated intact and translated in vitro when ribosomes and other necessary components are

added. They can also be 'reverse translated' in vitro by the viral enzyme reverse transcriptase into a cDNA molecule.

Most of the mRNA molecules (2/3 of them) of eukaryotes contain at their 3' end a poly(A) tail. The poly(A) tail is not coded in the DNA, but is added to the RNA in the nucleus after transcription. The addition of poly(A) is catalysed by the enzyme poly(A)polymerase, which recognizes the free 3'OH end of the mRNA and adds some 200 A residues. In some cases the presence of poly(A) does seem to affect the stability of the mRNA; no other effects have yet been found.The presence of a poly(A) tail makes it possible to prepare in one single step a relatively pure sample of mRNA. This is accomplished by attachment of an oligo(T) nucleotide on paramagnetic beads or on any other solid support. This oligo(T) is then used to anneal with the mRNA so that non-poly(A) containing RNA can be removed.

The eukaryotic mRNA has a methylated cap at its 5' end. The transcription starts normally with a purine (A or G). The first nucleotide retains its 5' triphosphate group, and makes the usual phosphodiester bond from its 3' position to the 5' position of the next nucleotide. A nuclear enzyme, guanylyl transferase, catalyses immediately after the onset of the transcription the addition of a terminal G, connected by an unusual 5'-5' triphosphate linkage.This structure is called the cap of the mRNA. It is a substrate for methylation at certain positions that occur in a specific order. The first methylation occurs in all eukaryotes, and consists of the addition of a methyl group to the 7 position of the terminal guanine. The enzyme responsible for this, guanine-7-methyl-transferase, is present in the cytoplasm. The resulting cap is referred to as cap0. The next step is the addition of another methyl group, to the 2'-O position of the penultimate base. This reaction is catalyzed by another enzyme (2'-O-methyl-transferase). A cap containing two methyl groups is called cap1. It is the predominant form of the cap in all eukaryotes except for the unicellular organisms where the reaction stops at cap0. In a small minority of cases in higher eukaryotes, another methyl group can be added by 2'-O-methyladenosine-N6-methyltransferase to this same base when it concerns an adenine. In some species, a methyl group may be added to the second base of the original transcript. The substrate for this reaction is the cap1 mRNA that already possesses two methyl groups. It creates the cap2 mRNA and constitutes some 10 to 15 % of the total capped population.

Transfer RNA

The transfer RNA or tRNA provides the 'adapter' with the twin functions of being used to recognize both the codon and the amino acid. The tRNA was first identified as a fraction of RNA sedimenting at 4S. Typically, tRNA's are 75-85 nucleotides in length. The transfer RNA molecules contain many of the 'unusual' and a vast range of modified bases. These modifications confer on tRNA a much greater range of structural versatility, which presumably is important for its various functions. Some of the modifications occur in positions involved in base pairing and therefore influence the stability of pairing.

The nucleotide sequence of every tRNA can be written out in the form of a cloverleaf, in which complementary base pairing forms the stems for single stranded loops. The stem-loop structures are known as the arms of tRNA. However, the actual form of the tRNA molecule is very compact. It folds in an L shaped tertiary structure with the amino acid linkage at one end and the anticodon at the other extremity. In addition to the base pairing that is represented in the secondary structure, further H bonds form in the tertiary structure. Some of the bases are involved in triple pairing.

Ribosomal RNA

Ribosomes, the protein synthesizing apparatus constitute a major component (20,000 ribosomes/genome). They contain about 10 % of the total bacterial protein and account for 80 % or so of the total mass of cellular RNA. A pointer to the importance of rRNA is that its sequence remains almost invariant within a cell, although it is coded by many genes. Two percent of the ribosomal RNA nucleotides are methylated. The great majority (80 %) of the methyl groups are carried on the 2' position of the ribose, the remainder are modifications of the bases, mostly adenines.The large (**50S**) subunit of the *E. coli* ribosome contains 34 proteins and two rRNAs: **5S** and **23S**. The **23S** rRNA is the real catalytic agent in peptide bond formation. Both rRNA molecules adopt a compact base-paired structure. One protein (the L7=L12 protein) is present in 4 copies -- all others are present as a single copy. One protein (L26) is the same as a protein (S20) in the small subunit.

The large (**60S**) subunit of eukaryotic ribosomes contains approximately 50 proteins and 3 rRNA molecules: **28S**, **5.8S** and **5S**. The **28S** and **5.8S** rRNA are both related to the bacterial **23S** rRNA. The **5.8S** rRNA is similar in sequence to the 5' end of the **23S** rRNA so its existence is probably a result of some ancient mutation that divided the ancestral gene in two.

The crystal structure of the ribosome complex (or its subunits) is solved. The left part of figure 73a shows the structure of the 23S rRNA (grey) from *Haloarcula marismortui*. Ribosomal proteins are shown in yellow. The catalytic centre is right where the star shines. This perspective is that from the 30S subunit.

Solving the structure of the ribosome clarified many issues on their function. A major function of the rRNA clearly is structural. Proteins bind to it at particular sites and in a specific order that is required for assembly of the subunit.

The 3' terminal region of bacterial 16S rRNA is highly conserved. It has two features that may be significant in protein synthesis. First, it is self-complementary, and could form a base-paired hairpin. Second, it contains the sequence CCUCCU. In all but one of the *E.coli* mRNA initiation sites, there is a sequence complementary to at least a trinucleotide part of this, and more usually to a length of 4 to 5 bases. Thus bacterial mRNA contains part or all of the oligonucleotide AGGAGG. This polypurine stretch is often known as the Shine-Dalgarno box. It lays 4 to 7 bases before the AUG initiation codon.

In prokaryotic systems, recognition of the initiation codon is based on the presence of a **Shine-Dalgarno** sequence, **complementary to the 3'-end of 16S ribosomal RNA** in the small ribosomal subunit (30S). The initiation codon is about 10 nt downstream. This mechanism allows for multiple initiation sites in a polycistronic mRNA.

The sequence at the 3' end of rRNA is remarkably well conserved between prokayotes and eukaryotes. Surprisingly the Shine Dalgarno box has been deleted in eukaryotes. This means that eukaryotic ribosomes should not be able to initiate properly the translation on bacterial mRNA molecules. Within the highly conserved 3' terminal region there is a highly purine-rich sequence, UGCGGAAGGAU, that could participate in hairpin formation. Some eukaryotic mRNA's have a tetra or pentanucleotide sequence able in theory to pair with part of

this (if the hairpin were disrupted). Hence, in eukaryotic systems, there is no corresponding sequence complementarity between mRNA and the corresponding 18S rRNA, and initiation is based on the 5'-cap site. The initiation codon is separated from the cap by a variable length **5'-UTR** or **untranslated region**, a minimum of 15 nt, and commonly 50-100 nt. Very few instances of true polycistronic mRNA have been reported for eukaryotic genes (but see later comments on upstream ORFs), although there are instances of regulated selection of alternative initiation sites from different candidate AUG codons. Normal ribosome entry is dependent on the 5'-cap, so reinitiation at an internal site on mRNA is rare, limiting most eukaryotic mRNAs to a single coding sequence, unlike polycistronic mRNA in prokaryotes.Certain viral mRNAs, e.g. encephalomyocarditis virus (EMCV), initiate at an internal AUG site by a **non-cap, non-scanning mechanism** (**IRES** initiation). In addition to the two major rRNA's the large subunit also contains a molecule of 5S RNA. All 5S RNA molecules display a highly base-paired structure, although there has been some difficulty in settling on an exact model.

guide RNA

The discovery of Blum et al. (Cell **60**, 189, 1990), of a class of RNA's known as guide RNA (gRNA) solved the intriguing observation of RNA editing. These small RNA's are encoded both in the maxicircle and minicircle DNA of kinetoplasts, and their 5' ends exhibit extensive sequence complementarity with the region that is 3' to the editing sites of cryptogenic transcripts.

Telomerase RNA

When DNA polymerase copies a DNA molecule, it requires not only a template to copy but a primer from which to initiate polymerisation. Thus, if the 3' end of a DNA strand is annealed to the primer it will not be copied by the polymerase and a linear DNA will tend to become shorter each time it is copied. In eukaryotes this problem is solved by telomeres. An enzyme, telomerase, capable of generating additional repeat units in compensation for loss during replication has been characterized. It was demonstrated that telomerase is a ribonucleoprotein. The RNA component has been suggested to function as the template for the addition of the correct GGGGTT units to the chromosome end. Indeed, the RNA of the telomerase contained the sequence CAACCCCAA that is complementary to one and a half of the

telomeric units. Furthermore, mutations in the CAACCCCAA sequence caused the synthesis, i*n vivo*, of modified telomeric sequences that were complementary to the mutated sequences.

ORGANIZATION OF THE GENETIC MATERIAL IN CHROMOSOMES

Within the cell, DNA is associated with proteins. Each DNA molecule and its associated protein molecules is called a chromosome. This is valid for all kinds of organisms, viruses, prokaryotes and eukaryotes. The various proteins in chromatin perform a number of essential functions:

- -(i) numerous DNA-binding proteins catalyse and regulate vital cellular processes and DNA transactions such as DNA replication, transcription, DNA modification (methylation …), repair of DNA damage, DNA recombination, DNA translocation ...
- -(ii) DNA-binding proteins compact the DNA (histones in eukaryotes, small basic histone-like proteins in bacteria)
- -(iii) DNA binding proteins protect the DNA from degradation and chemical damage (naked DNA is much more rapidly degraded by nucleases and more sensitive to oxidative damage)

Only approximately 3% of the nucleoid is DNA.

Prokaryotes and eukaryotes show important differences in the organization of their genetic material. **Prokaryotes (Bacteria and Archaea) have no nucleus**. Their DNA is not surrounded by a nuclear membrane and appears as a granular structure associated with the membrane, the nucleoid. Mostly, the genetic material of a prokaryotic cell consists of one **single circular DNA molecule** ranging from 1.55 Mbp (1,550 kbp, *Picrophilus torridus* an acidophilic archaeon, currently the smallest genome of a free living organism, not a parasite) to approximately 9.1 Mbp (*Myxococcus xanthus*, a δ-proteobacterium). Symbionts and parasites such as *Mycoplasma* and *Nanoarchaeum,* have an even smaller genome of only 0.5 Mbp. *Streptomyces lividans* has a single linear chromosome. Some bacteria have more than one chromosome: for example, *Sinorhizobium meliloti* has three circular chromosomes and *Agrobacterium tumefaciens* has four chromosomes (3 circular + 1linear). When prokaryotic cells are dividing rapidly, portions of the chromosome in the process of replication are present in 2 or sometimes even four copies. Prokaryotes also frequently carry one or more smaller independent circular DNAs called plasmids or episomes. Plasmids do not integrate into the main chromosome, episomes can reside in the cell as independent molecules or canintegrate into the main chromosome. Plasmids and

episomes are generally not essential for bacterial growth. They carry genes that confer desirable traits to the bacteria, such as antibiotic resistance, allow the transfer of genetic information from one cell to another by means of conjugation (such as the F episome or fertility factor of *Escherichia coli*), or carry virulence factors. The *Escherichia coli* genome is approximately 4.7 Mbp long. Its sequence has been entirely determined in 1997. The dimensions of an *E. coli* cell are only 1.5 x 2 to 6 μm. Therefore, the DNA has to be strongly compacted to fit into the cell otherwise it would have a diameter of 430 μm, or a length of about 1 mm). This compaction of the chromosome is realized by small basic proteins that bind to the DNA, mostly in a non-sequence-specific manner (sequence-independent). They organize the genome into compacted loops such that the genome is divided into domains (dynamic organization in about 400 independent supercoiled domains of 10 kb each for the *E. coli* chromosome). These proteins that help in the compaction of bacterial and archaeal genomes are frequently referred to as histone-like proteins [IHF (Integration Host Factor), H-NS (histone-like nucleoid structuring protein), HU (heat-unstable nucleoid protein), FIS (factor for inversion stimulation), etc). This name is however somewhat misleading since these proteins are totally different from the eukaryotic histones. The compacted *E. coli* nucleoid occupies approximately 15% of the cellular volume. Some archaea, the euryarchaeota, have real equivalents of the eukaryotic histone heteroteramer (H3-H4)2 that show the same characteristic histone fold. The crenarchaeota have no equivalents of the eukaryotic histones. They compact their DNA by the use of small basic proteins (Sul7, Alba, Cren7), more similar to the bacterial way of compacting.

Eukaryotes are characterized by a nuclear membrane that surrounds their genetic patrimonium. The nucleus is an organelle with a diameter of several μm and is mostly visible in the light microscope. The membrane has some 3,000 to 4,000 pores of 9 nm diameter, which allow the passage of macromolecules up to 60,000 Dalton (Da) and contains numerous proteins involved in active transport of small and macromolecules in (proteins, cDNA) and out (mature mRNA) of the nucleus. DNA replication and transcription take place in the nucleus, but protein synthesis occurs in the cytoplasm. Transcription and translation of mRNA will, therefore, take place in different cellular compartments and will be uncoupled in space and in time. The nuclear membrane plays a very important role in the transport of RNA (inside →outside) and proteins (outside → inside). **The genetic material of eukaryotes is organized in several**

chromosomes. In eukaryotic chromatin the mass of protein and the mass of DNA are more or less equal. Chromatin contains several types of proteins among which SMC (Structural Maintenance of the Chromosome), topoisomerases, transcription regulators and histones (largest part). Histone proteins are highly conserved among eukaryotic cells. The histone proteins H3 and H4 are nearly identical in all eukaryotes (only 2 out of 102 residues differ between H4 of peas and cows, 8 differences between yeast and humans). **Eukaryotic chromosomes are linear molecules** (unlike the vast majority of prokaryotic chromosomes that are circular), except for the mitochondrial and chloroplast DNA's, that are circular.

Mitochondrial genomes are rather small circular molecules, about 16.5 kb in humans and 78 kb in *S. cerevisiae*. Chromosome means colored body. This reflects the fact that chromosomes were first discovered in the light microscope by using staining techniques. They are mainly invisible except at the moment of the cell division, when they are compacted and condensed. In eukaryotes the **DNA is tightly associated with basic proteins, the histones,** which assure the compaction by wrapping of the DNA around histone octamers (see below, Chapter II. 6.). These condensed structures are called nucleosomes; one nucleosome contains about 200 bp. The succession of **nucleosomes** forms a fibrous structure called **chromatin** (10 nm fiber). Chromatin can be further condensed by folding and bending to form a 30 nm fiber, which in turn is arranged as loops around a proteinaceous scaffold to form a chromosome. **The majority of the eukaryotic cells are diploid**. They contain two copies of each chromosome (except the sex chromosome). The two copies are called **homologs**; one is derived from each parent. A subset of the eukaryotic cells is either haploid or polyploid. Haploid cells contain a single copy of each chromosome and are for instance involved in the sexual reproduction. Eggs and spermatozoids are haploid cells. Yeast is also a haploid organism for most of its cell cycle. **Polyploid cell have more than two copies of each chromosome.** Polyploidy is mainly associated with plant cells. Some other organisms maintain the majority of their adult cells in a polyploid state. In extreme cases the number of copies can be as high as 100 or 1,000.

This type of chromosome amplification allows the cell to generate larger amounts of mRNA and thus protein. For example, megakaryocytes are specialized polyploid cells (□128 sets of chromosomes) that produce thousands of platelets that lack chromosomes but are essential

components of human blood (200,000 platelets/ml of blood). The segregation of large numbers of chromosomes is very difficult, therefore polyploid cells have almost always stopped dividing. The diploid human genome (ensemble of all the chromosomes) would be about 2 m long without compaction. The complete genetic information is stored in 46 chromosomes with a total length of only 200 μm. This indicates a compaction of about 10,000-fold. Strong compaction by several orders of magnitude is a must because the nucleus of a human cell is only 10-15 μm in diameter. **The total amount of DNA in the haploid genome is called the C-value**. There is an enormous variation in the range of C-values of different organisms: from <106 bp for a mycoplasma to > 1011 bp for some plants and amphibians. It appears that the minimal amount of DNA required for a member of the different evolutionary phyla (the smallest genome size found for a member of each group) increases from prokaryotes to eukaryotes.The DNA content of a nucleus and the number of chromosomes can vary largely from one organism to another one.

- *Giardia* (protozoan) 12 Mbp 4 chromosomes
- *Saccharomyces cerevisiae* (baker's yeast) 12 Mbp 16 chromosomes, haploid
-*Schizosaccharomyces pombe* (fission yeast) 12 Mbp 3 chromosomes, haploid
- *Arabidopsis thaliana* (weed) 125 Mbp 5 pairs, diploid
- *Drosophila melanogaster* (fruit fly) 180 Mbp 4 pairs, diploid
- *Homo sapiens* 4,800 Mbp 23 pairs, diploid

Many plants and amphibians have even more DNA per cell than humans.The smallest human chromosome contains about 10-fold more bps than the *E. coli* genome. The haploid form of the human genome has about 200-fold more DNA/cell than yeast, but it has only a few chromosomes more. Therefore, the DNA content of chromosomes can vary widely.It appears that genome size is roughly correlated with the complexity of the organism. Prokaryotic cells typically have genomes of less than 10 Mbp and many Archaea have rather small genomes compised between 2 and 3 Mbps. The genomes of single cell eukaryotes are typically less than 50 Mbp (more complex protozoans can have up to 200 Mbp genomes).

Multi cellular eukaryotes have even larger genomes, up to greater than 100,000 Mbp. Nevertheless, organisms of apparently similar complexity can have very different genome sizes. A fruit fly has a genome 25-fold smaller than a locust, and rice has a genome that is

approximately 40-fold smaller than that of wheat. This lack of correlation between genome size and genetic complexity is referred to as the **C-value paradox**. It is presently not understood why natural selection allows this variation and whether it has evolutionary consequences. The fruit fly *Drosophila melanogaster* (genome = 180 Mbp) has about 15,000 genes that code for proteins (180 genes/Mbp). A human cell (4,800 Mbp) has about 20,000 to 25,000 of such genes (9.3 genes/Mbp). Only approximately 3% of the human genome is used as information coding for proteins. The remaining 97% are essentially used to regulate the expression of the other 3%. It consists of introns, repetitive DNA, snRNA genes (small nuclear RNA), etc. So, its is clearly not the number of translated genes that determines the more complex behavior and superior possibilities of humans, but rather the way in which this information is expressed and used, and the number of connections that can be made among gene products (alternative splicing greatly enhances the number of different proteins that can be obtained from a single gene). Clearly, the **genome density** (number of genes/Mbp of DNA) can vary largely and it appears that more complex organisms have a much lower gene density. Therefore, different organisms use the gene-encoding potential of DNA with varying efficiencies. The highest gene densities are found in viruses. In some instances they use both strands of a given DNA region to encode overlapping genes. In bacteria overlapping genes are quite rare but the gene density is still high, about 1,000 genes/ Mbp (on average 1 gene per kbp). The gene density of *S. cerevisiae* is about 500 genes/ Mbp (about half of the gene density of prokaryotes). But the human genome has a gene density that is still 50-fold lower (9.3 genes/Mbp). Two factors contribute to the reduction in gene density:

(i) an increase in gene size due to the presence of introns

(ii) an increase in the length of intergenic sequences

These intergenic regions consist of repetitive DNA: microsatellites (< 13 bp in length, highly repetitive DNA with many thousand copies per genome, frequently organized as long tandem repeats), genome wide repeats (> 100 bp up to > 1 kb) and pseudo genes. Although there are numerous classes of repeats, their common feature is that they are all forms of transposable elements (or remnants there off, inactive forms of transposable elements). The percentage of repeated DNA varies widely. Prokaryotes have nearly no repetitive DNA sequences, in lower eukaryotes it represents about 20 % of the genome, in animal cells it may represent up to 50 % of the genome (about 40% in humans) and in some plant cells and amphibians it can represent even

up to 80%, so that the nonrepetitive DNA is reduced to a minority. Although it is common to refer to repeated DNA as junk DNA, the stable maintenance of these sequences over hundreds to thousands of generations suggests that intergenic DNA confers a positive value or selective advantage to the host organisms. Obviously, much has still to be discovered in this domain.

Structure of eukaryotic chromosomes

Different levels of eukaryotic chromosome structure and organization can be observed in the microscope. Long before it was clear that chromosomes are the source of genetic information in the cell, their movements and changes during cell division were already well analyzed; the compact nature of condensed mitotic chromosomes allows their easy observation in the light microscope. In the electron microscope two other states of chromatin could be readily observed: the 30-nm and the 10-nm fibers. Indeed, chromosomal DNA in the interphase is much less compact. The 30-nm fiber appears as a structure folded into large loops reaching out of a protein scaffold. The 10-nm fiber is the least compact form of chromatin. It resembles a regular series of "beads on a string". The beads are the nucleosomes. They are the building blocks of eukaryotic chromosomes.

Structure of the nucleosome

A nucleosome is composed of a core of eight histone proteins and the DNA wrapped around this core in a left-handed manner. This wrapping introduces negative supercoiling (toroidal supercoiling) in the eukaryotic DNA. Wrapped DNA has about 10.2 bp/turn instead of 10.5 in non-wrapped DNA. It are the positively charged N-terminal tails of the histones and the way they protrude from the histone core that impose the direction of the wrapping. The DNA between the nucleosomes is called **linker DNA**. By assembling into nucleosomes the DNA is compacted about 6-fold (much less than the 10,000-fold observed in chromosomes). The DNA that is tightly associated with the histones is called the **core DNA**. It is □147 bp long and is wound 1.65-times around the outside of the histone octamer, like thread around a spool. This is so in all eukaryotic cells. In contrast, the length of the linker DNA is variable (in *S. cerevisiae* it is on average 13-18 bp, in humans 38-53 bp long). In any cell there are also stretches of DNA

that are not packaged into nucleosomes. These are typically regions that are being replicated or transcribed.These sites are then typically associated with non-histone proteins that are either involved in these processes or are regulating these processes. The accessibility of the DNA and the chromatin remodeling are important aspects of eukaryotic transcription and transcription regulation. Histones are small, positively charged proteins (they contain more than 20% lysine and arginine residues). They are by far the most abundant class of DNA-associated proteins. Eukaryotic cells generally contain five different abundant histones: H1, H2A, H2B, H3 and H4. The histones H2A, H2B, H3 and H4 are the core histones. They are 11 to 15 kDa proteins that form the octameric disc shaped core around which the DNA is wrapped. Histone H1 (20 kDa) binds to the linker DNA and is referred to as linker histone. H1 is half as abundant as the core histones (which are present in equal amounts). The core histones assemble in an ordered fashion only in the presence of DNA. In solution they form intermediate assemblies. A conserved region, called the histonefold domain (a very characteristic fold) mediates the assembly of the histone-only intermediates. The histone fold is composed of three α-helical regions separated by two short unstructured loops. The histone fold mediates the formation of head to tail heterodimers of specific pairs of histones. First H3 and H4 form heterodimers that then assemble into a tetramer (H3-H4)2. H2A and H2B form heterodimers (not tetramers). The further ordered assembly of the nucleosome involves the association of these building blocks with DNA. First the (H3-H4)2 tetramer binds to DNA. Then two H2A-H2B dimers join the (H3-H4)2 tetramer - DNA complex to form the final nucleosome. The assembly of nucleosomes on replicating DNA requires the CAF-I complex (Chromatin Assembly Complex I). The four core histones have an N-terminal extension, called the tails. The tails lacks defined structure and are accessible within the intact nucleosome (treatment with trypsin which specifically cleaves proteins after positively-charged residues readily removes the N-terminal tails). These tails are the sites of extensive modifications that alter the function of individual nucleosomes. These modifications mainly include: phosphorylation of serine residues and acetylation and methylation of lysine and arginine residues. These modifications alter the interaction of the DNA with the histone core and play a major role in modification of chromatin structure and therefore affect processes such as transcription initiation (see Eukaryotic transcription). Histone tail modifications also alter the molecular interactions between adjacent nucleosomes, thereby changing the level of chromatin compaction. Furthermore, eukaryotic cells have several variants of the histone proteins H2A, H3

and H1 that confer special properties to the chromatin structure. Fourteen distinct sites of contact can be observed between the histones and the core nucleosomal DNA. That is one for each time the minor groove of the DNA faces the histone octamer. This generates about 140 hydrogen bonds between the histone proteins and DNA. The majority of the hydrogen bonds are between the proteins and the oxygen atoms in the phosphodiester backbone near the minor groove of the DNA (non sequence specific interactions). Only 7 hydrogen bonds are formed between the proteins side chains and the base-specific groups in the minor groove (none of these is with elements that distinguish between a G-C and an A-T bp). This high number of hydrogen bonds (most protein-DNA interactions will make only ± 20 hydrogen bonds) provides the driving force to bend the DNA around the histone core. The highly positive charge of the histone proteins also serves to mask the negative charge of the phosphates on the inside of the bend into unfavorable close proximity and

facilitates the close juxtaposition of the two α-helices in the 1.65-times wrapped structure. The organization of DNA into nucleosomes is dynamic. Three forms of mobility can be observed:

(i) sliding of the histone octamer along the DNA

(ii) complete transfer of the histone octamer from one molecule to another

(iii) more subtle remodeling of the protein-DNA interactions within a nucleosome.

Higher-order chromatin structure: binding of H1 stabilizes 30 nm fibers

The next step in the packaging of DNA, once the nucleosomes are formed, is the binding of the linker histone H1. The basic protein H1 makes contacts with two distinct regions: it binds the linker DNA and the DNA helix in the middle of the nucleosome-bound DNA, thereby further tightening the histone-DNA association and bringing these two regions in close proximity. Therefore, H1 binding increases the length of the DNA wrapped tightly around the histone-octamer. Actively transcribed regions generally have no H1 bound. There are two models for the 30-nm fiber. In the solenoid model, the nucleosomal DNA forms a superhelix containing ± 6 nucleosomes per turn. In this model, the flat surfaces on either face of the histone octamer disc are adjacent to each other and the DNA surface of the nucleosomes forms the outside, accessible surface of the superhelix. The linker DNA is buried in the center of the superhelix, but it never passes through the axis of the fiber. It circles around the central axis as the DNA moves from one nucleosome to the next.

An alternative model for the 30-nm fiber is the **zig-zag model**. In this zig-zag model, the linker DNA is required to pass through the central axis of the fiber in a relatively straight form. Longer linker DNA will favor this conformation. Because the average length of the linker DNA varies between different species, the form of the 30-nm fiber may not always be the same. It has been observed that core histones lacking the N-terminal tails are incapable of forming 30-nm fibers. Therefore, the N-terminal tails are absolutely required and their most likely role is to stabilize the 30-nm fiber structure by interacting with adjacent nucleosomes. The formation of the 30-nm fiber structure results in the compaction of the linear length of DNA by ± 40-fold (still insufficient). Additional folding of the 30-nm fiber is required to fit 1-2 meters of DNA in a nucleus of about 10-5 m (10 μm) across. The exact nature of this folding is still unclear, but it appears that the 30-nm fiber forms loops of 40-90 kb that are hold together at their bases by a **proteinaceous structure**, the **nuclear scaffold**. Topoisomerase II (Topo II, a type II topoisomerase) and the SMC proteins (Structural Maintenance Chromosome) are abundant components of the nuclear scaffold. These proteins are key components of the machinery that condenses and holds daughter chromosomes together after chromosome duplication.