= DNA =

Deoxyribonucleic acid (/ di??ksi?ra?bo?nj??kli??k , -?kle??k / ; DNA) is a molecule that carries the genetic instructions used in the growth , development , functioning and reproduction of all known living organisms and many viruses . DNA and RNA are nucleic acids ; alongside proteins and complex carbohydrates (polysaccharides) , they are one of the three major types of macromolecule that are essential for all known forms of life . Most DNA molecules consist of two biopolymer strands coiled around each other to form a double helix .

The two DNA strands are known as polynucleotides since they are composed of simpler units called nucleotides . Each nucleotide is composed of a nitrogen @-@ containing nucleobase ? either cytosine (C) , guanine (G) , adenine (A) , or thymine (T) ? as well as a sugar called deoxyribose and a phosphate group . The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next , resulting in an alternating sugar @-@ phosphate backbone . According to base pairing rules (A with T , and C with G) , hydrogen bonds bind the nitrogenous bases of the two separate polynucleotide strands to make double @-@ stranded DNA . The total amount of related DNA base pairs on Earth is estimated at 5 @.@ 0 x 1037 , and weighs 50 billion tonnes . In comparison , the total mass of the biosphere has been estimated to be as much as 4 TtC (trillion tons of carbon) .

DNA stores biological information . The DNA backbone is resistant to cleavage , and both strands of the double @-@ stranded structure store the same biological information . Biological information is replicated as the two strands are separated . A significant portion of DNA (more than 98 % for humans) is non @-@ coding , meaning that these sections do not serve as patterns for protein sequences .

The two strands of DNA run in opposite directions to each other and are therefore anti @-@ parallel . Attached to each sugar is one of four types of nucleobases (informally , bases) . It is the sequence of these four nucleobases along the backbone that encodes biological information . Under the genetic code , RNA strands are translated to specify the sequence of amino acids within proteins . These RNA strands are initially created using DNA strands as a template in a process called transcription .

Within eukaryotic cells , DNA is organized into long structures called chromosomes . During cell division these chromosomes are duplicated in the process of DNA replication , providing each cell its own complete set of chromosomes . Eukaryotic organisms (animals , plants , fungi , and protists) store most of their DNA inside the cell nucleus and some of their DNA in organelles , such as mitochondria or chloroplasts . In contrast , prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm . Within the eukaryotic chromosomes , chromatin proteins such as histones compact and organize DNA . These compact structures guide the interactions between DNA and other proteins , helping control which parts of the DNA are transcribed .

DNA was first isolated by Friedrich Miescher in 1869 . Its molecular structure was identified by James Watson and Francis Crick in 1953 , whose model @-@ building efforts were guided by X @-@ ray diffraction data acquired by Rosalind Franklin . DNA is used by researchers as a molecular tool to explore physical laws and theories , such as the ergodic theorem and the theory of elasticity . The unique material properties of DNA have made it an attractive molecule for material scientists and engineers interested in micro- and nano @-@ fabrication . Among notable advances in this field are DNA origami and DNA @-@ based hybrid materials .

= = Properties = =

DNA is a long polymer made from repeating units called nucleotides . The structure of DNA is non @-@ static , all species comprises two helical chains each coiled round the same axis , and each with a pitch of 34 ångströms (3 @.@ 4 nanometres) and a radius of 10 ångströms (1 @.@ 0 nanometre) . According to another study , when measured in a particular solution , the DNA chain measured 22 to 26 ångströms wide (2 @.@ 2 to 2 @.@ 6 nanometres) , and one nucleotide unit measured 3 @.@ 3 Å (0 @.@ 33 nm) long . Although each individual repeating unit is very small , DNA polymers can be very large molecules containing millions of nucleotides . For instance , the DNA in the largest human chromosome , chromosome number 1 , consists of approximately 220 million base pairs and would be 85 mm long if straightened .

In living organisms DNA does not usually exist as a single molecule , but instead as a pair of molecules that are held tightly together . These two long strands entwine like vines , in the shape of a double helix . The nucleotide repeats contain both the segment of the backbone of the molecule , which holds the chain together , and a nucleobase , which interacts with the other DNA strand in the helix . A nucleobase linked to a sugar is called a nucleoside and a base linked to a sugar and one or more phosphate groups is called a nucleotide . A polymer comprising multiple linked nucleotides (as in DNA) is called a polynucleotide .

The backbone of the DNA strand is made from alternating phosphate and sugar residues . The sugar in DNA is 2 @-@ deoxyribose , which is a pentose (five @-@ carbon) sugar . The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings . These asymmetric bonds mean a strand of DNA has a direction . In a double helix the direction of the nucleotides in one strand is opposite to their direction in the other strand : the strands are antiparallel . The asymmetric ends of DNA strands are called the 5 ? (five prime) and 3 ? (three prime) ends , with the 5 ? end having a terminal phosphate group and the 3 ? end a terminal hydroxyl group . One major difference between DNA and RNA is the sugar , with the 2 @-@ deoxyribose in DNA being replaced by the alternative pentose sugar ribose in RNA .

The DNA double helix is stabilized primarily by two forces : hydrogen bonds between nucleotides and base @-@ stacking interactions among aromatic nucleobases . In the aqueous environment of the cell , the conjugated ? bonds of nucleotide bases align perpendicular to the axis of the DNA molecule , minimizing their interaction with the solvation shell and therefore , the Gibbs free energy . The four bases found in DNA are adenine (abbreviated A) , cytosine (C) , guanine (G) and thymine (T) . These four bases are attached to the sugar / phosphate to form the complete nucleotide , as shown for adenosine monophosphate . Adenine pairs with thymine and guanine pairs with cytosine . It was represented by A @-@ T base pairs and G @-@ C base pairs .

= = = Nucleobase classification = = =

The nucleobases are classified into two types : the purines , A and G , being fused five- and six @-@ membered heterocyclic compounds , and the pyrimidines , the six @-@ membered rings C and T. A fifth pyrimidine nucleobase , uracil (U) , usually takes the place of thymine in RNA and differs from thymine by lacking a methyl group on its ring . In addition to RNA and DNA a large number of artificial nucleic acid analogues have also been created to study the properties of nucleic acids , or for use in biotechnology .

Uracil is not usually found in DNA , occurring only as a breakdown product of cytosine . However , in a number of bacteriophages ? Bacillus subtilis bacteriophages PBS1 and PBS2 and Yersinia bacteriophage piR1 @-@ 37 ? thymine has been replaced by uracil . Another phage - Staphylococcal phage S6 - has been identified with a genome where thymine has been replaced by uracil .

Base J (beta @-@ d @-@ glucopyranosyloxymethyluracil) , a modified form of uracil , is also found in a number of organisms : the flagellates Diplonema and Euglena , and all the kinetoplastid genera . Biosynthesis of J occurs in two steps : in the first step a specific thymidine in DNA is converted into hydroxymethyldeoxyuridine ; in the second HOMedU is glycosylated to form J. Proteins that bind specifically to this base have been identified . These proteins appear to be distant relatives of the Tet1 oncogene that is involved in the pathogenesis of acute myeloid leukemia . J appears to act as a termination signal for RNA polymerase II .

= = = Grooves = = =

Twin helical strands form the DNA backbone . Another double helix may be found tracing the spaces , or grooves , between the strands . These voids are adjacent to the base pairs and may provide a binding site . As the strands are not symmetrically located with respect to each other , the grooves are unequally sized . One groove , the major groove , is 22 Å wide and the other , the minor groove , is 12 Å wide . The width of the major groove means that the edges of the bases are more accessible in the major groove than in the minor groove . As a result , proteins such as transcription factors that can bind to specific sequences in double @-@ stranded DNA usually make contact with the sides of the bases exposed in the major groove . This situation varies in unusual conformations of DNA within the cell (see below) , but the major and minor grooves are always named to reflect the differences in size that would be seen if the DNA is twisted back into the ordinary B form .

= = = Base pairing = = =

In a DNA double helix , each type of nucleobase on one strand bonds with just one type of nucleobase on the other strand . This is called complementary base pairing . Here , purines form hydrogen bonds to pyrimidines , with adenine bonding only to thymine in two hydrogen bonds , and cytosine bonding only to guanine in three hydrogen bonds . This arrangement of two nucleotides binding together across the double helix is called a base pair . As hydrogen bonds are not covalent , they can be broken and rejoined relatively easily . The two strands of DNA in a double helix can therefore be pulled apart like a zipper , either by a mechanical force or high temperature . As a result of this complementarity , all the information in the double @-@ stranded sequence of a DNA helix is duplicated on each strand , which is vital in DNA replication . Indeed , this reversible and specific interaction between complementary base pairs is critical for all the functions of DNA in living organisms .

The two types of base pairs form different numbers of hydrogen bonds , AT forming two hydrogen bonds , and GC forming three hydrogen bonds (see figures , right) . DNA with high GC @-@ content is more stable than DNA with low GC @-@ content .

As noted above , most DNA molecules are actually two polymer strands , bound together in a helical fashion by noncovalent bonds ; this double stranded structure (dsDNA) is maintained largely by the intrastrand base stacking interactions , which are strongest for G , C stacks . The two strands can come apart ? a process known as melting ? to form two single @-@ stranded DNA molecules (ssDNA) molecules . Melting occurs at high temperature , low salt and high pH (low pH also melts DNA , but since DNA is unstable due to acid depurination , low pH is rarely used) .

The stability of the dsDNA form depends not only on the GC @-@ content (% G , C basepairs) but also on sequence (since stacking is sequence specific) and also length (longer molecules are more stable) . The stability can be measured in various ways ; a common way is the " melting temperature " , which is the temperature at which 50 % of the ds molecules are converted to ss molecules ; melting temperature is dependent on ionic strength and the concentration of DNA . As a result , it is both the percentage of GC base pairs and the overall length of a DNA double helix that determines the strength of the association between the two strands of DNA . Long DNA helices with a high GC @-@ content have stronger @-@ interacting strands , while short helices with high AT content have weaker @-@ interacting strands . In biology , parts of the DNA double helix that need to separate easily , such as the TATAAT Pribnow box in some promoters , tend to have a high AT content , making the strands easier to pull apart .

In the laboratory , the strength of this interaction can be measured by finding the temperature necessary to break the hydrogen bonds , their melting temperature (also called Tm value) . When all the base pairs in a DNA double helix melt , the strands separate and exist in solution as two entirely independent molecules . These single @-@ stranded DNA molecules (ssDNA) have no single common shape , but some conformations are more stable than others .

= = = Sense and antisense = = =

A DNA sequence is called " sense " if its sequence is the same as that of a messenger RNA copy that is translated into protein . The sequence on the opposite strand is called the " antisense " sequence . Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands can contain both sense and antisense sequences) . In both prokaryotes and eukaryotes , antisense RNA sequences are produced , but the functions of these RNAs are not entirely clear . One proposal is that antisense RNAs are involved in regulating gene expression through RNA @-@ RNA base pairing .

A few DNA sequences in prokaryotes and eukaryotes , and more in plasmids and viruses , blur the distinction between sense and antisense strands by having overlapping genes . In these cases , some DNA sequences do double duty , encoding one protein when read along one strand , and a second protein when read in the opposite direction along the other strand . In bacteria , this overlap may be involved in the regulation of gene transcription , while in viruses , overlapping genes increase the amount of information that can be encoded within the small viral genome .

= = = Supercoiling = = =

DNA can be twisted like a rope in a process called DNA supercoiling . With DNA in its " relaxed " state , a strand usually circles the axis of the double helix once every 10 @.@ 4 base pairs , but if the DNA is twisted the strands become more tightly or more loosely wound . If the DNA is twisted in the direction of the helix , this is positive supercoiling , and the bases are held more tightly together . If they are twisted in the opposite direction , this is negative supercoiling , and the bases come apart more easily . In nature , most DNA has slight negative supercoiling that is introduced by enzymes called topoisomerases . These enzymes are also needed to relieve the twisting stresses introduced into DNA strands during processes such as transcription and DNA replication .

= = = Alternative DNA structures = = =

DNA exists in many possible conformations that include A @-@ DNA , B @-@ DNA , and Z @-@ DNA forms , although , only B @-@ DNA and Z @-@ DNA have been directly observed in functional organisms . The conformation that DNA adopts depends on the hydration level , DNA sequence , the amount and direction of supercoiling , chemical modifications of the bases , the type and concentration of metal ions , as well as the presence of polyamines in solution .

The first published reports of A @-@ DNA X @-@ ray diffraction patterns ? and also B @-@ DNA ? used analyses based on Patterson transforms that provided only a limited amount of structural information for oriented fibers of DNA . An alternative analysis was then proposed by Wilkins et al . , in 1953 , for the in vivo B @-@ DNA X @-@ ray diffraction / scattering patterns of highly hydrated DNA fibers in terms of squares of Bessel functions . In the same journal , James Watson and Francis Crick presented their molecular modeling analysis of the DNA X @-@ ray diffraction patterns to suggest that the structure was a double @-@ helix .

Although the " B @-@ DNA form " is most common under the conditions found in cells , it is not a well @-@ defined conformation but a family of related DNA conformations that occur at the high hydration levels present in living cells . Their corresponding X @-@ ray diffraction and scattering patterns are characteristic of molecular paracrystals with a significant degree of disorder .

Compared to B @-@ DNA , the A @-@ DNA form is a wider right @-@ handed spiral , with a shallow , wide minor groove and a narrower , deeper major groove . The A form occurs under non @-@ physiological conditions in partially dehydrated samples of DNA , while in the cell it may be produced in hybrid pairings of DNA and RNA strands , as well as in enzyme @-@ DNA complexes . Segments of DNA where the bases have been chemically modified by methylation may undergo a larger change in conformation and adopt the Z form . Here , the strands turn about the helical axis in a left @-@ handed spiral , the opposite of the more common B form . These unusual structures can be recognized by specific Z @-@ DNA binding proteins and may be involved in the regulation of transcription .

= = = Alternative DNA chemistry = = =

For a number of years exobiologists have proposed the existence of a shadow biosphere , a postulated microbial biosphere of Earth that uses radically different biochemical and molecular processes than currently known life . One of the proposals was the existence of lifeforms that use arsenic instead of phosphorus in DNA . A report in 2010 of the possibility in the bacterium GFAJ @-@ 1 , was announced , though the research was disputed , and evidence suggests the bacterium actively prevents the incorporation of arsenic into the DNA backbone and other biomolecules .

= = = Quadruplex structures = = =

At the ends of the linear chromosomes are specialized regions of DNA called telomeres . The main function of these regions is to allow the cell to replicate chromosome ends using the enzyme telomerase , as the enzymes that normally replicate DNA cannot copy the extreme 3 ? ends of chromosomes . These specialized chromosome caps also help protect the DNA ends , and stop the DNA repair systems in the cell from treating them as damage to be corrected . In human cells , telomeres are usually lengths of single @-@ stranded DNA containing several thousand repeats of a simple TTAGGG sequence .

These guanine @-@ rich sequences may stabilize chromosome ends by forming structures of stacked sets of four @-@ base units , rather than the usual base pairs found in other DNA molecules . Here , four guanine bases form a flat plate and these flat four @-@ base units then stack on top of each other , to form a stable G @-@ quadruplex structure . These structures are stabilized by hydrogen bonding between the edges of the bases and chelation of a metal ion in the centre of each four @-@ base unit . Other structures can also be formed , with the central set of four bases coming from either a single strand folded around the bases , or several different parallel strands , each contributing one base to the central structure .

In addition to these stacked structures , telomeres also form large loop structures called telomere loops , or T @-@ loops . Here , the single @-@ stranded DNA curls around in a long circle stabilized by telomere @-@ binding proteins . At the very end of the T @-@ loop , the single @-@ stranded telomere DNA is held onto a region of double @-@ stranded DNA by the telomere strand disrupting the double @-@ helical DNA and base pairing to one of the two strands . This triple @-@ stranded structure is called a displacement loop or D @-@ loop .

= = = Branched DNA = = =

In DNA , fraying occurs when non @-@ complementary regions exist at the end of an otherwise complementary double @-@ strand of DNA . However , branched DNA can occur if a third strand of DNA is introduced and contains adjoining regions able to hybridize with the frayed regions of the pre @-@ existing double @-@ strand . Although the simplest example of branched DNA involves only three strands of DNA , complexes involving additional strands and multiple branches are also possible . Branched DNA can be used in nanotechnology to construct geometric shapes , see the section on uses in technology below .

= = Chemical modifications and altered DNA packaging = =

= = = Base modifications and DNA packaging = = =

The expression of genes is influenced by how the DNA is packaged in chromosomes , in a structure called chromatin . Base modifications can be involved in packaging , with regions that have low or no gene expression usually containing high levels of methylation of cytosine bases . DNA packaging and its influence on gene expression can also occur by covalent modifications of the histone protein core around which DNA is wrapped in the chromatin structure or else by remodeling carried out by chromatin remodeling complexes (see Chromatin remodeling) . There is , further , crosstalk between DNA methylation and histone modification , so they can coordinately affect chromatin and gene expression .

For one example , cytosine methylation , produces 5 @-@ methylcytosine , which is important for X @-@ chromosome inactivation . The average level of methylation varies between organisms ? the worm Caenorhabditis elegans lacks cytosine methylation , while vertebrates have higher levels , with up to 1 % of their DNA containing 5 @-@ methylcytosine . Despite the importance of 5 @-@ methylcytosine , it can deaminate to leave a thymine base , so methylated cytosines are particularly prone to mutations . Other base modifications include adenine methylation in bacteria , the presence of 5 @-@ hydroxymethylcytosine in the brain , and the glycosylation of uracil to produce the " J @-@ base " in kinetoplastids .

= = = Damage = = =

DNA can be damaged by many sorts of mutagens , which change the DNA sequence . Mutagens include oxidizing agents , alkylating agents and also high @-@ energy electromagnetic radiation such as ultraviolet light and X @-@ rays . The type of DNA damage produced depends on the type of mutagen . For example , UV light can damage DNA by producing thymine dimers , which are cross @-@ links between pyrimidine bases . On the other hand , oxidants such as free radicals or hydrogen peroxide produce multiple forms of damage , including base modifications , particularly of guanosine , and double @-@ strand breaks . A typical human cell contains about 150 @,@ 000 bases that have suffered oxidative damage . Of these oxidative lesions , the most dangerous are double @-@ strand breaks , as these are difficult to repair and can produce point mutations , insertions and deletions from the DNA sequence , as well as chromosomal translocations . These mutations can cause cancer . Because of inherent limitations in the DNA repair mechanisms , if humans lived long enough , they would all eventually develop cancer . DNA damages that are naturally occurring , due to normal cellular processes that produce reactive oxygen species , the hydrolytic activities of cellular water , etc . , also occur frequently . Although most of these damages are repaired , in any cell some DNA damage may remain despite the action of repair processes . These remaining DNA damages accumulate with age in mammalian postmitotic tissues . This accumulation appears to be an important underlying cause of aging .

Many mutagens fit into the space between two adjacent base pairs , this is called intercalation . Most intercalators are aromatic and planar molecules ; examples include ethidium bromide , acridines , daunomycin , and doxorubicin . For an intercalator to fit between base pairs , the bases must separate , distorting the DNA strands by unwinding of the double helix . This inhibits both transcription and DNA replication , causing toxicity and mutations . As a result , DNA intercalators may be carcinogens , and in the case of thalidomide , a teratogen . Others such as benzo [a] pyrene diol epoxide and aflatoxin form DNA adducts that induce errors in replication . Nevertheless , due to their ability to inhibit DNA transcription and replication , other similar toxins are also used in chemotherapy to inhibit rapidly growing cancer cells .

= = Biological functions = =

DNA usually occurs as linear chromosomes in eukaryotes , and circular chromosomes in prokaryotes . The set of chromosomes in a cell makes up its genome ; the human genome has approximately 3 billion base pairs of DNA arranged into 46 chromosomes . The information carried by DNA is held in the sequence of pieces of DNA called genes . Transmission of genetic information in genes is achieved via complementary base pairing . For example , in transcription , when a cell uses the information in a gene , the DNA sequence is copied into a complementary RNA sequence through the attraction between the DNA and the correct RNA nucleotides . Usually , this RNA copy is then used to make a matching protein sequence in a process called translation , which depends on the same interaction between RNA nucleotides . In alternative fashion , a cell may simply copy its genetic information in a process called DNA replication . The details of these functions are covered in other articles ; here the focus is on the interactions between DNA and other molecules that mediate the function of the genome .

= = = Genes and genomes = = =

Genomic DNA is tightly and orderly packed in the process called DNA condensation to fit the small available volumes of the cell . In eukaryotes , DNA is located in the cell nucleus , as well as small amounts in mitochondria and chloroplasts . In prokaryotes , the DNA is held within an irregularly shaped body in the cytoplasm called the nucleoid . The genetic information in a genome is held within genes , and the complete set of this information in an organism is called its genotype . A gene is a unit of heredity and is a region of DNA that influences a particular characteristic in an organism . Genes contain an open reading frame that can be transcribed , as well as regulatory sequences such as promoters and enhancers , which control the transcription of the open reading frame .

In many species , only a small fraction of the total sequence of the genome encodes protein . For example , only about 1 @.@ 5 % of the human genome consists of protein @-@ coding exons , with over 50 % of human DNA consisting of non @-@ coding repetitive sequences . The reasons for the presence of so much noncoding DNA in eukaryotic genomes and the extraordinary differences in genome size , or C @-@ value , among species represent a long @-@ standing puzzle known as the " C @-@ value enigma " . However , some DNA sequences that do not code protein may still encode functional non @-@ coding RNA molecules , which are involved in the regulation of gene expression .

Some noncoding DNA sequences play structural roles in chromosomes . Telomeres and centromeres typically contain few genes , but are important for the function and stability of chromosomes . An abundant form of noncoding DNA in humans are pseudogenes , which are copies of genes that have been disabled by mutation . These sequences are usually just molecular fossils , although they can occasionally serve as raw genetic material for the creation of new genes through the process of gene duplication and divergence .

= = = Transcription and translation = = =

A gene is a sequence of DNA that contains genetic information and can influence the phenotype of an organism . Within a gene , the sequence of bases along a DNA strand defines a messenger RNA sequence , which then defines one or more protein sequences . The relationship between the nucleotide sequences of genes and the amino @-@ acid sequences of proteins is determined by the rules of translation , known collectively as the genetic code . The genetic code consists of three @-@ letter ' words ' called codons formed from a sequence of three nucleotides (e.g. ACT , CAG , TTT) .

In transcription , the codons of a gene are copied into messenger RNA by RNA polymerase . This RNA copy is then decoded by a ribosome that reads the RNA sequence by base @-@ pairing the messenger RNA to transfer RNA , which carries amino acids . Since there are 4 bases in 3 @-@ letter combinations , there are 64 possible codons (43 combinations) . These encode the twenty standard amino acids , giving most amino acids more than one possible codon . There are also three ' stop ' or ' nonsense ' codons signifying the end of the coding region ; these are the TAA , TGA , and TAG codons .

= = = Replication = = =

Cell division is essential for an organism to grow , but , when a cell divides , it must replicate the DNA in its genome so that the two daughter cells have the same genetic information as their parent . The double @-@ stranded structure of DNA provides a simple mechanism for DNA replication . Here , the two strands are separated and then each strand 's complementary DNA sequence is recreated by an enzyme called DNA polymerase . This enzyme makes the complementary strand by finding the correct base through complementary base pairing , and bonding it onto the original strand . As DNA polymerases can only extend a DNA strand in a 5 ? to 3 ? direction , different mechanisms are used to copy the antiparallel strands of the double helix . In this way , the base on the old strand dictates which base appears on the new strand , and the cell ends up with a perfect copy of its DNA .

= = = Extracellular nucleic acids = = =

Naked extracellular DNA (eDNA) , most of it released by cell death , is nearly ubiquitous in the environment . Its concentration in soil may be as high as 2 ?g / L , and its concentration in natural aquatic environments may be as high at 88 ?g / L. Various possible functions have been proposed for eDNA : it may be involved in horizontal gene transfer ; it may provide nutrients ; and it may act as a buffer to recruit or titrate ions or antibiotics . Extracellular DNA acts as a functional extracellular matrix component in the biofilms of a number of bacterial species . It may act as a recognition factor to regulate the attachment and dispersal of specific cell types in the biofilm ; it may contribute to biofilm formation ; and it may contribute to the biofilm 's physical strength and resistance to biological stress .

= = Interactions with proteins = =

All the functions of DNA depend on interactions with proteins . These protein interactions can be non @-@ specific , or the protein can bind specifically to a single DNA sequence . Enzymes can also bind to DNA and of these , the polymerases that copy the DNA base sequence in transcription and DNA replication are particularly important .

= = = DNA @-@ binding proteins = = =

Structural proteins that bind DNA are well @-@ understood examples of non @-@ specific DNA @-@ protein interactions . Within chromosomes , DNA is held in complexes with structural proteins . These proteins organize the DNA into a compact structure called chromatin . In eukaryotes this structure involves DNA binding to a complex of small basic proteins called histones , while in prokaryotes multiple types of proteins are involved . The histones form a disk @-@ shaped complex called a nucleosome , which contains two complete turns of double @-@ stranded DNA wrapped around its surface . These non @-@ specific interactions are formed through basic residues in the histones making ionic bonds to the acidic sugar @-@ phosphate backbone of the DNA , and are therefore largely independent of the base sequence . Chemical modifications of these basic amino acid residues include methylation , phosphorylation and acetylation . These chemical changes alter the strength of the interaction between the DNA and the histones , making the DNA more or less accessible to transcription factors and changing the rate of transcription . Other non @-@ specific DNA @-@ binding proteins in chromatin include the high @-@ mobility group proteins , which bind to bent or distorted DNA . These proteins are important in bending arrays of nucleosomes and arranging them into the larger structures that make up chromosomes .

A distinct group of DNA @-@ binding proteins are the DNA @-@ binding proteins that specifically bind single @-@ stranded DNA . In humans , replication protein A is the best @-@ understood member of this family and is used in processes where the double helix is separated , including DNA replication , recombination and DNA repair . These binding proteins seem to stabilize single @-@ stranded DNA and protect it from forming stem @-@ loops or being degraded by nucleases .

In contrast , other proteins have evolved to bind to particular DNA sequences . The most intensively studied of these are the various transcription factors , which are proteins that regulate transcription . Each transcription factor binds to one particular set of DNA sequences and activates or inhibits the transcription of genes that have these sequences close to their promoters . The transcription factors do this in two ways . Firstly , they can bind the RNA polymerase responsible for transcription , either directly or through other mediator proteins ; this locates the polymerase at the promoter and allows it to begin transcription . Alternatively , transcription factors can bind enzymes that modify the histones at the promoter . This changes the accessibility of the DNA template to the polymerase .

As these DNA targets can occur throughout an organism 's genome , changes in the activity of one type of transcription factor can affect thousands of genes . Consequently , these proteins are often the targets of the signal transduction processes that control responses to environmental changes or cellular differentiation and development . The specificity of these transcription factors ' interactions with DNA come from the proteins making multiple contacts to the edges of the DNA bases , allowing them to " read " the DNA sequence . Most of these base @-@ interactions are made in the major groove , where the bases are most accessible .

= = = DNA @-@ modifying enzymes = = =

= = = = Nucleases and ligases = = = =

Nucleases are enzymes that cut DNA strands by catalyzing the hydrolysis of the phosphodiester bonds . Nucleases that hydrolyse nucleotides from the ends of DNA strands are called exonucleases , while endonucleases cut within strands . The most frequently used nucleases in molecular biology are the restriction endonucleases , which cut DNA at specific sequences . For instance , the EcoRV enzyme shown to the left recognizes the 6 @-@ base sequence 5 ? -GATATC @-@ 3 ? and makes a cut at the vertical line . In nature , these enzymes protect bacteria against phage infection by digesting the phage DNA when it enters the bacterial cell , acting as part of the restriction modification system . In technology , these sequence @-@ specific nucleases are used in molecular cloning and DNA fingerprinting .

Enzymes called DNA ligases can rejoin cut or broken DNA strands . Ligases are particularly important in lagging strand DNA replication , as they join together the short segments of DNA produced at the replication fork into a complete copy of the DNA template . They are also used in DNA repair and genetic recombination .

= = = = Topoisomerases and helicases = = = =

Topoisomerases are enzymes with both nuclease and ligase activity . These proteins change the amount of supercoiling in DNA . Some of these enzymes work by cutting the DNA helix and allowing one section to rotate , thereby reducing its level of supercoiling ; the enzyme then seals the DNA break . Other types of these enzymes are capable of cutting one DNA helix and then passing a second strand of DNA through this break , before rejoining the helix . Topoisomerases are required for many processes involving DNA , such as DNA replication and transcription .

Helicases are proteins that are a type of molecular motor . They use the chemical energy in nucleoside triphosphates , predominantly ATP , to break hydrogen bonds between bases and unwind the DNA double helix into single strands . These enzymes are essential for most processes where enzymes need to access the DNA bases .

= = = = Polymerases = = = =

Polymerases are enzymes that synthesize polynucleotide chains from nucleoside triphosphates . The sequence of their products are created based on existing polynucleotide chains ? which are called templates . These enzymes function by repeatedly adding a nucleotide to the 3 ? hydroxyl group at the end of the growing polynucleotide chain . As a consequence , all polymerases work in a 5 ? to 3 ? direction . In the active site of these enzymes , the incoming nucleoside triphosphate base @-@ pairs to the template : this allows polymerases to accurately synthesize the complementary strand of their template . Polymerases are classified according to the type of template that they use .

In DNA replication , DNA @-@ dependent DNA polymerases make copies of DNA polynucleotide chains . In order to preserve biological information , it is essential that the sequence of bases in each copy are precisely complementary to the sequence of bases in the template strand . Many DNA polymerases have a proofreading activity . Here , the polymerase recognizes the occasional mistakes in the synthesis reaction by the lack of base pairing between the mismatched nucleotides . If a mismatch is detected , a 3 ? to 5 ? exonuclease activity is activated and the incorrect base removed . In most organisms , DNA polymerases function in a large complex called the replisome that contains multiple accessory subunits , such as the DNA clamp or helicases .

RNA @-@ dependent DNA polymerases are a specialized class of polymerases that copy the sequence of an RNA strand into DNA . They include reverse transcriptase , which is a viral enzyme involved in the infection of cells by retroviruses , and telomerase , which is required for the replication of telomeres . Telomerase is an unusual polymerase because it contains its own RNA template as part of its structure .

Transcription is carried out by a DNA @-@ dependent RNA polymerase that copies the sequence of a DNA strand into RNA . To begin transcribing a gene , the RNA polymerase binds to a sequence of DNA called a promoter and separates the DNA strands . It then copies the gene sequence into a messenger RNA transcript until it reaches a region of DNA called the terminator , where it halts and detaches from the DNA . As with human DNA @-@ dependent DNA polymerases , RNA polymerase II , the enzyme that transcribes most of the genes in the human genome , operates as part of a large protein complex with multiple regulatory and accessory subunits .

= = Genetic recombination = =

A DNA helix usually does not interact with other segments of DNA , and in human cells the different chromosomes even occupy separate areas in the nucleus called " chromosome territories " . This physical separation of different chromosomes is important for the ability of DNA to function as a stable repository for information , as one of the few times chromosomes interact is in chromosomal crossover which occurs during sexual reproduction , when genetic recombination occurs . Chromosomal crossover is when two DNA helices break , swap a section and then rejoin .

Recombination allows chromosomes to exchange genetic information and produces new combinations of genes , which increases the efficiency of natural selection and can be important in the rapid evolution of new proteins . Genetic recombination can also be involved in DNA repair , particularly in the cell 's response to double @-@ strand breaks .

The most common form of chromosomal crossover is homologous recombination , where the two chromosomes involved share very similar sequences . Non @-@ homologous recombination can be damaging to cells , as it can produce chromosomal translocations and genetic abnormalities . The recombination reaction is catalyzed by enzymes known as recombinases , such as RAD51 . The first step in recombination is a double @-@ stranded break caused by either an endonuclease or damage to the DNA . A series of steps catalyzed in part by the recombinase then leads to joining of the two helices by at least one Holliday junction , in which a segment of a single strand in each helix is annealed to the complementary strand in the other helix . The Holliday junction is a tetrahedral junction structure that can be moved along the pair of chromosomes , swapping one strand for another . The recombination reaction is then halted by cleavage of the junction and re @-@ ligation of the released DNA .

= = Evolution = =

DNA contains the genetic information that allows all modern living things to function , grow and reproduce . However , it is unclear how long in the 4 @-@ billion @-@ year history of life DNA has performed this function , as it has been proposed that the earliest forms of life may have used RNA as their genetic material . RNA may have acted as the central part of early cell metabolism as it can both transmit genetic information and carry out catalysis as part of ribozymes . This ancient RNA world where nucleic acid would have been used for both catalysis and genetics may have influenced the evolution of the current genetic code based on four nucleotide bases . This would occur , since the number of different bases in such an organism is a trade @-@ off between a small number of bases increasing replication accuracy and a large number of bases increasing the catalytic efficiency of ribozymes . However , there is no direct evidence of ancient genetic systems , as recovery of DNA from most fossils is impossible because DNA survives in the environment for less than one million years , and slowly degrades into short fragments in solution . Claims for older DNA have been made , most notably a report of the isolation of a viable bacterium from a salt crystal 250 million years old , but these claims are controversial .

Building blocks of DNA (adenine , guanine and related organic molecules) may have been formed extraterrestrially in outer space . Complex DNA and RNA organic compounds of life , including uracil , cytosine and thymine , have also been formed in the laboratory under conditions mimicking those found in outer space , using starting chemicals , such as pyrimidine , found in meteorites . Pyrimidine , like polycyclic aromatic hydrocarbons (PAHs) , the most carbon @-@ rich chemical found in the universe , may have been formed in red giants or in interstellar dust and gas clouds .

= = Uses in technology = =

= = = Genetic engineering = = =

Methods have been developed to purify DNA from organisms , such as phenol @-@ chloroform extraction , and to manipulate it in the laboratory , such as restriction digests and the polymerase chain reaction . Modern biology and biochemistry make intensive use of these techniques in recombinant DNA technology . Recombinant DNA is a man @-@ made DNA sequence that has been assembled from other DNA sequences . They can be transformed into organisms in the form of plasmids or in the appropriate format , by using a viral vector . The genetically modified organisms produced can be used to produce products such as recombinant proteins , used in medical research , or be grown in agriculture .

= = = DNA profiling = = =

Forensic scientists can use DNA in blood , semen , skin , saliva or hair found at a crime scene to identify a matching DNA of an individual , such as a perpetrator . This process is formally termed DNA profiling , but may also be called " genetic fingerprinting " . In DNA profiling , the lengths of variable sections of repetitive DNA , such as short tandem repeats and minisatellites , are compared between people . This method is usually an extremely reliable technique for identifying a matching DNA . However , identification can be complicated if the scene is contaminated with DNA from several people . DNA profiling was developed in 1984 by British geneticist Sir Alec Jeffreys , and first used in forensic science to convict Colin Pitchfork in the 1988 Enderby murders case .

The development of forensic science , and the ability to now obtain genetic matching on minute samples of blood , skin , saliva or hair has led to a re @-@ examination of a number of cases . Evidence can now be uncovered that was not scientifically possible at the time of the original examination . Combined with the removal of the double jeopardy law in some places , this can allow cases to be reopened where previous trials have failed to produce sufficient evidence to convince a jury . People charged with serious crimes may be required to provide a sample of DNA for matching purposes . The most obvious defence to DNA matches obtained forensically is to claim that cross @-@ contamination of evidence has taken place . This has resulted in meticulous strict handling procedures with new cases of serious crime . DNA profiling is also used to identify victims of mass casualty incidents . As well as positively identifying bodies or body parts in serious accidents , DNA profiling is being successfully used to identify individual victims in mass war graves ? matching to family members .

DNA profiling is also used in DNA paternity testing in order to determine if someone is the biologicalparent or grandparent of a child with the probability of parentage is typically 99 @.@ 99 % when the alleged parent is biologically related to the child . Normal DNA sequencing methods happen after birth but there are new methods to test paternity while the mother is still pregnant .

= = = DNA enzymes or catalytic DNA = = =

Deoxyribozymes , also called DNAzymes or catalytic DNA are first discovered in 1994 . They are mostly single stranded DNA sequences isolated from a large pool of random DNA sequences through a combinatorial approach called in vitro selection or SELEX . DNAzymes catalyze variety of chemical reactions including RNA / DNA cleavage , RNA / DNA ligation , amino acids phosphorylation / dephosphorylation , carbon @-@ carbon bond formation , and etc . DNAzymes can enhance catalytic rate of chemical reactions up to 100 @,@ 000 @,@ 000 @,@ 000 @-@ fold over the uncatalyzed reaction . The most extensively studied class of DNAzymes are RNA @-@ cleaving DNAzymes which have been used in detection of different metal ions and designing therapeutic agents . Several metal @-@ specific DNAzymes have been reported including the GR @-@ 5 DNAzyme (lead @-@ specific) , the CA1 @-@ 3 DNAzymes (copper @-@ specific) , the 39E DNAzyme (uranyl @-@ specific) and the NaA43 DNAzyme (sodium @-@ specific) . The NaA43 DNAzyme , which is reported to be more than 10 @,@ 000 @-@ fold selective for sodium over other metal ions , was used to make a real @-@ time sodium sensor in living cells .

= = = Bioinformatics = = =

Bioinformatics involves the development of techniques to store , data mine , search and manipulate biological data , including DNA nucleic acid sequence data . These have led to widely applied advances in computer science , especially string searching algorithms , machine learning and database theory . String searching or matching algorithms , which find an occurrence of a sequence of letters inside a larger sequence of letters , were developed to search for specific sequences of nucleotides . The DNA sequence may be aligned with other DNA sequences to identify homologous sequences and locate the specific mutations that make them distinct . These techniques , especially multiple sequence alignment , are used in studying phylogenetic relationships and protein function . Data sets representing entire genomes ' worth of DNA sequences , such as those produced by the Human Genome Project , are difficult to use without the annotations that identify the locations of genes and regulatory elements on each chromosome . Regions of DNA sequence that have the characteristic patterns associated with protein- or RNA @-@ coding genes can be identified by gene finding algorithms , which allow researchers to predict the presence of particular gene products and their possible functions in an organism even before they have been isolated experimentally . Entire genomes may also be compared , which can shed light on the evolutionary history of particular organism and permit the examination of complex evolutionary events .

= = = DNA nanotechnology = = =

DNA nanotechnology uses the unique molecular recognition properties of DNA and other nucleic acids to create self @-@ assembling branched DNA complexes with useful properties . DNA is thus used as a structural material rather than as a carrier of biological information . This has led to the creation of two @-@ dimensional periodic lattices (both tile @-@ based and using the " DNA origami " method) as well as three @-@ dimensional structures in the shapes of polyhedra . Nanomechanical devices and algorithmic self @-@ assembly have also been demonstrated , and these DNA structures have been used to template the arrangement of other molecules such as gold nanoparticles and streptavidin proteins .

= = = History and anthropology = = =

Because DNA collects mutations over time , which are then inherited , it contains historical information , and , by comparing DNA sequences , geneticists can infer the evolutionary history of organisms , their phylogeny . This field of phylogenetics is a powerful tool in evolutionary biology . If DNA sequences within a species are compared , population geneticists can learn the history of particular populations . This can be used in studies ranging from ecological genetics to anthropology ; For example , DNA evidence is being used to try to identify the Ten Lost Tribes of Israel .

= = = Information storage = = =

In a paper published in Nature in January 2013 , scientists from the European Bioinformatics Institute and Agilent Technologies proposed a mechanism to use DNA 's ability to code information as a means of digital data storage . The group was able to encode 739 kilobytes of data into DNA code , synthesize the actual DNA , then sequence the DNA and decode the information back to its original form , with a reported 100 % accuracy . The encoded information consisted of text files and audio files . A prior experiment was published in August 2012 . It was conducted by researchers at Harvard University , where the text of a 54 @,@ 000 @-@ word book was encoded in DNA .

= = History of DNA research = =

DNA was first isolated by the Swiss physician Friedrich Miescher who , in 1869 , discovered a microscopic substance in the pus of discarded surgical bandages . As it resided in the nuclei of cells , he called it " nuclein " . In 1878 , Albrecht Kossel isolated the non @-@ protein component of " nuclein " , nucleic acid , and later isolated its five primary nucleobases . In 1919 , Phoebus Levene identified the base , sugar and phosphate nucleotide unit . Levene suggested that DNA consisted of a string of nucleotide units linked together through the phosphate groups . Levene thought the chain was short and the bases repeated in a fixed order . In 1937 , William Astbury produced the first X @-@ ray diffraction patterns that showed that DNA had a regular structure .

In 1927 , Nikolai Koltsov proposed that inherited traits would be inherited via a " giant hereditary molecule " made up of " two mirror strands that would replicate in a semi @-@ conservative fashion using each strand as a template " . In 1928 , Frederick Griffith in his experiment discovered that traits of the " smooth " form of Pneumococcus could be transferred to the " rough " form of the same bacteria by mixing killed " smooth " bacteria with the live " rough " form . This system provided the first clear suggestion that DNA carries genetic information ? the Avery ? MacLeod ? McCarty experiment ? when Oswald Avery , along with coworkers Colin MacLeod and Maclyn McCarty , identified DNA as the transforming principle in 1943 . DNA 's role in heredity was confirmed in 1952 , when Alfred Hershey and Martha Chase in the Hershey ? Chase experiment showed that DNA is the genetic material of the T2 phage .

In 1953 , James Watson and Francis Crick suggested what is now accepted as the first correct double @-@ helix model of DNA structure in the journal Nature . Their double @-@ helix , molecular model of DNA was then based on a single X @-@ ray diffraction image (labeled as " Photo 51 ") taken by Rosalind Franklin and Raymond Gosling in May 1952 , as well as the information that the DNA bases are paired .

Experimental evidence supporting the Watson and Crick model was published in a series of five articles in the same issue of Nature . Of these , Franklin and Gosling 's paper was the first publication of their own X @-@ ray diffraction data and original analysis method that partially supported the Watson and Crick model ; this issue also contained an article on DNA structure by Maurice Wilkins and two of his colleagues , whose analysis and in vivo B @-@ DNA X @-@ ray patterns also supported the presence in vivo of the double @-@ helical DNA configurations as proposed by Crick and Watson for their double @-@ helix molecular model of DNA in the previous two pages of Nature . In 1962 , after Franklin 's death , Watson , Crick , and Wilkins jointly received the Nobel Prize in Physiology or Medicine . Nobel Prizes are awarded only to living recipients . A debate continues about who should receive credit for the discovery .

In an influential presentation in 1957 , Crick laid out the central dogma of molecular biology , which foretold the relationship between DNA , RNA , and proteins , and articulated the " adaptor hypothesis " . Final confirmation of the replication mechanism that was implied by the double @-@ helical structure followed in 1958 through the Meselson ? Stahl experiment . Further work by Crick and coworkers showed that the genetic code was based on non @-@ overlapping triplets of bases , called codons , allowing Har Gobind Khorana , Robert W. Holley and Marshall Warren Nirenberg to decipher the genetic code . These findings represent the birth of molecular biology .