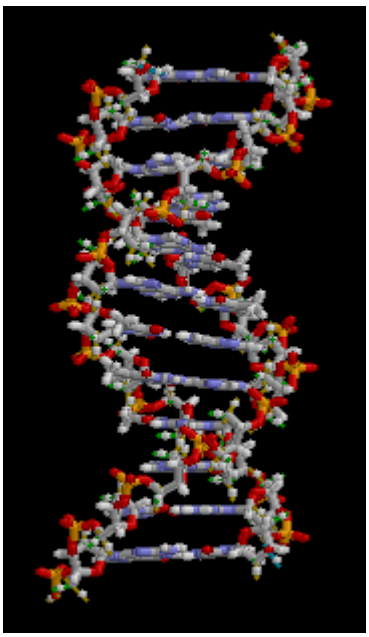
DNA

From Wikipedia, the free encyclopedia



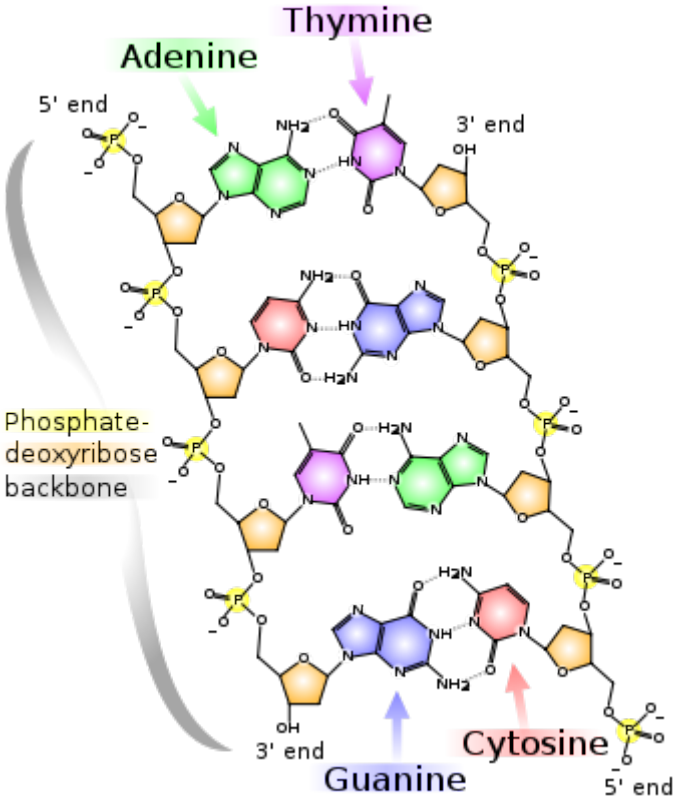
The structure of part of a DNA double helix

Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses. The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints, like a recipe or a code, since it contains the instructions needed to construct other components of cells, such as proteins and RNA molecules. The DNA segments that carry this genetic information are called genes, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information.

Chemically, DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are therefore anti-parallel. Attached to each sugar is one of four types of molecules called bases. It is the sequence of these four bases along the backbone that encodes information. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related nucleic acid RNA, in a process called transcription.

Within cells, DNA is organized into long structures called chromosomes. These chromosomes are duplicated before cells divide, in a process called DNA replication. 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.[1] In contrast, prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm. Within the 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.

Contents

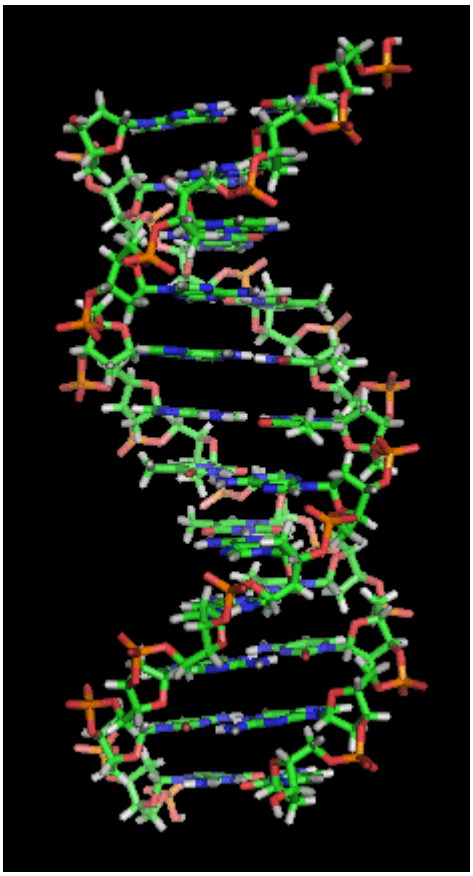


Chemical structure of DNA. Hydrogen bonds shown as dotted lines.

DNA is a long polymer made from repeating units called nucleotides.[2][3][4] The DNA chain is 22 to 26 Ångströms wide (2.2 to 2.6 nanometres), and one nucleotide unit is 3.3 Å (0.33 nm) long.[5] Although each individual repeating unit is very small, DNA polymers can be very large molecules containing millions of nucleotides. For instance, the largest human chromosome, chromosome number 1, is approximately 220 million base pairs long.[6]

In living organisms, DNA does not usually exist as a single molecule, but instead as a pair of molecules that are held tightly together.[7][8] 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 base, which interacts with the other DNA strand in the helix. A base 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. If multiple nucleotides are linked together, as in DNA, this polymer is called a polynucleotide.[9]

The backbone of the DNA strand is made from alternating phosphate and sugar residues.[10] 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.[8]



A section of DNA. The bases lie horizontally between the two spiraling strands.[11] Animated version at File:DNA orbit animated.gif.

The DNA double helix is stabilized by hydrogen bonds between the bases attached to the two strands. 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.

These bases are classified into two types; adenine and guanine are fused five- and six-membered heterocyclic compounds called purines, while cytosine and thymine are six-membered rings called pyrimidines.[8] A fifth pyrimidine base, called uracil (U), usually takes the place of thymine in RNA and differs from thymine by lacking a methyl group on its ring. Uracil is not usually found in DNA, occurring only as a breakdown product of cytosine. In addition to RNA and DNA, a large number of artificial nucleic acid analogues have also been created to study the proprieties of nucleic acids, or for use in biotechnology.[12]

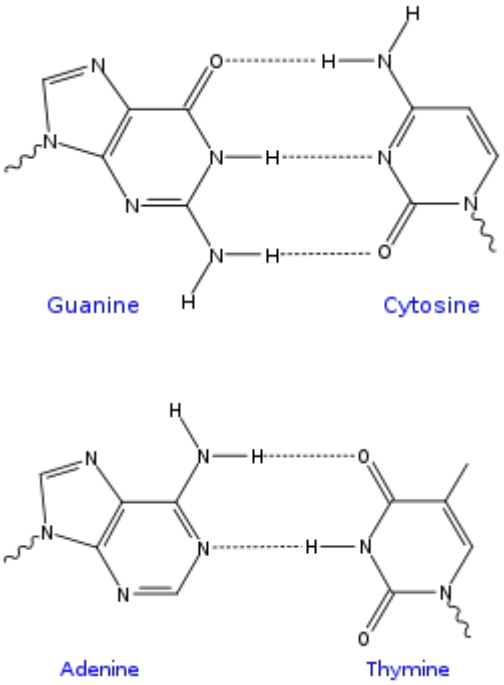
Grooves

Twin helical strands form the DNA backbone. Another double helix may be found by 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 directly opposite 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.[13] The narrowness of the minor groove means that the edges of the bases are more accessible in the major groove. As a result, proteins like transcription factors that can bind to specific sequences in double-stranded DNA usually make contacts to the sides of the bases exposed in the major groove.[14] 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

Each type of base on one strand forms a bond with just one type of base on the other strand. This is called complementary base pairing. Here, purines form hydrogen bonds to pyrimidines, with A bonding only to T, and C bonding only to G. 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.[15] 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.[3]



Top, a GC base pair with three hydrogen bonds. Bottom, an AT base pair with two hydrogen bonds. Non-covalent hydrogen bonds between the pairs are shown as dashed lines.

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, left). DNA with high GC-content is more stable than DNA with low GC-content, but contrary to popular belief, this is not due to the extra hydrogen bond of a GC base pair but rather the contribution of stacking interactions (hydrogen bonding merely provides specificity of the pairing, not stability).[16] As a result, it is both the percentage of GC base pairs and the overall length of a DNA double helix that determine 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.[17] 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.[18] In the laboratory, the strength of this interaction can be measured by finding the temperature required 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.[19]

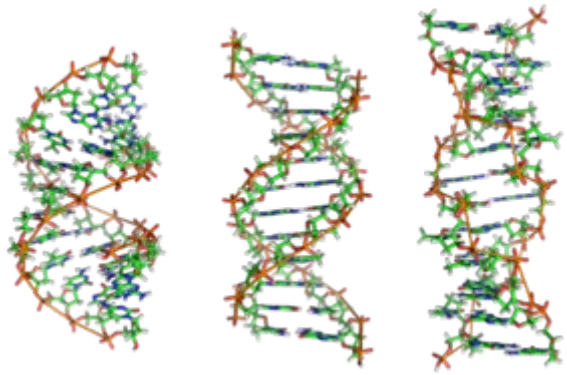
Sense and antisense

A DNA sequence is called "sense" if its sequence is the same as that of a messenger RNA copy that is translated into protein.[20] 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 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.[21] One proposal is that antisense RNAs are involved in regulating gene expression through RNA-RNA base pairing.[22]

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.[23] 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,[24] while in viruses, overlapping genes increase the amount of information that can be encoded within the small viral genome.[25]

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.[26] 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.[27] These enzymes are also needed to relieve the twisting stresses introduced into DNA strands during processes such as transcription and DNA replication.[28]



From left to right, the structures of A, B and Z DNA

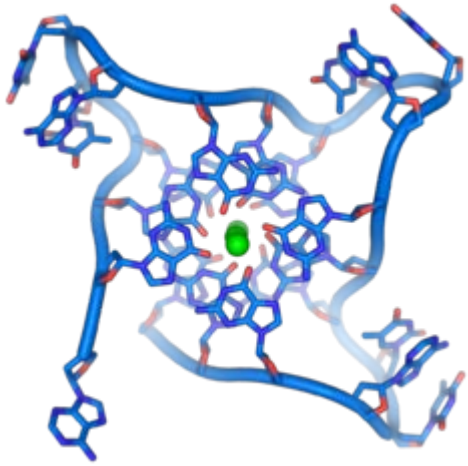
Alternate DNA structures

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.[10] 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.[29]

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.[30][31] An alternate 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.[32] In the same journal, James D. 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.[7] Although the `B-DNA form' is most common under the conditions found in cells,[33] it is not a well-defined conformation but a family of related DNA conformations[34] 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.[35][36]

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.[37][38] 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.[39] These unusual structures can be recognized by specific Z-DNA binding proteins and may be involved in the regulation of transcription.[40]

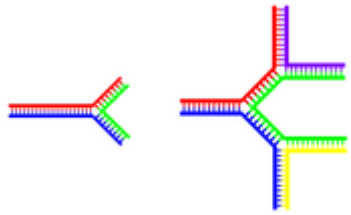


DNA quadruplex formed by telomere repeats. The looped conformation of the DNA backbone is very different from the typical DNA helix.[41]

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.[42] 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.[43] In human cells, telomeres are usually lengths of single-stranded DNA containing several thousand repeats of a simple TTAGGG sequence.[44] 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.[45] 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.[46] 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.[47] 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.[45]



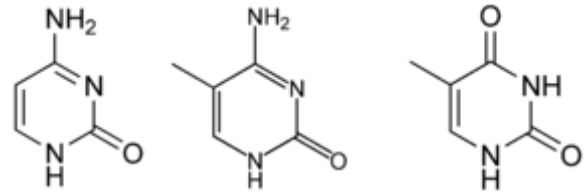
Single branch | Multiple branches

Branched DNA can form networks containing multiple branches.

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.[48] Branched DNA can be used in nanotechnology to construct geometric shapes, see the section on uses in technology below.
[edit] Vibration
Further information: Low-frequency collective motion in proteins and DNA
DNA may carry out low-frequency collective motion as observed by the Raman spectroscopy [49] [50] and analyzed with the quasi-continuum model. [51] [52]

Chemical modifications



cytosine | 5-methylcytosine | thymine

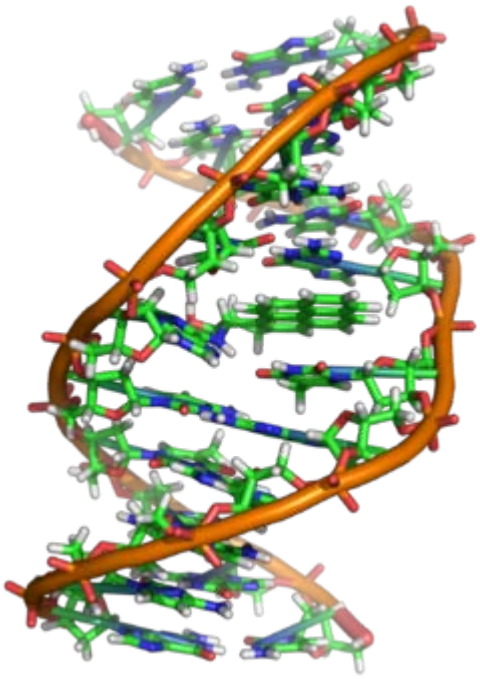
Structure of cytosine with and without the 5-methyl group. Deamination converts 5-methylcytosine into thymine.

Base modifications

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. For example, cytosine methylation, produces 5-methylcytosine, which is important for X-chromosome inactivation.[53] 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.[54] Despite the importance of 5-methylcytosine, it can deaminate to

leave a thymine base, so methylated cytosines are particularly prone to mutations.[55] Other base modifications include adenine methylation in bacteria, the presence of 5-hydroxymethylcytosine in the brain,[56] and the glycosylation of uracil to produce the "J-base" in kinetoplastids.[57][58]

Damage



A covalent adduct between a metabolically activated form of benzo[a]pyrene, the major mutagen in tobacco smoke, and DNA[59]

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.[60] 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.[61] A typical human cell contains about 150,000 bases that have suffered oxidative damage.[62] 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.[63]

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, daunomycin, and doxorubicin. In order 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 are often carcinogens, and benzo[a]pyrene diol epoxide, acridines, aflatoxin and ethidium bromide are well-known examples.[64][65][66] 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.[67]

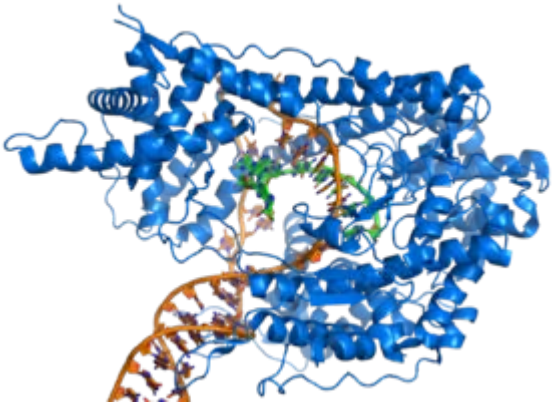
Biological functions

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.[68] 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. Alternatively, 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 we focus on the interactions between DNA and other molecules that mediate the function of the genome.

Genes and genomes

Genomic DNA is located in the cell nucleus of eukaryotes, 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.[69] 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.[70] The reasons for the presence of so much non-coding 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".[71] However, DNA sequences that do not code protein may still encode functional non-coding RNA molecules, which are involved in the regulation of gene expression.[72]



T7 RNA polymerase (blue) producing a mRNA (green) from a DNA template (orange).[73]

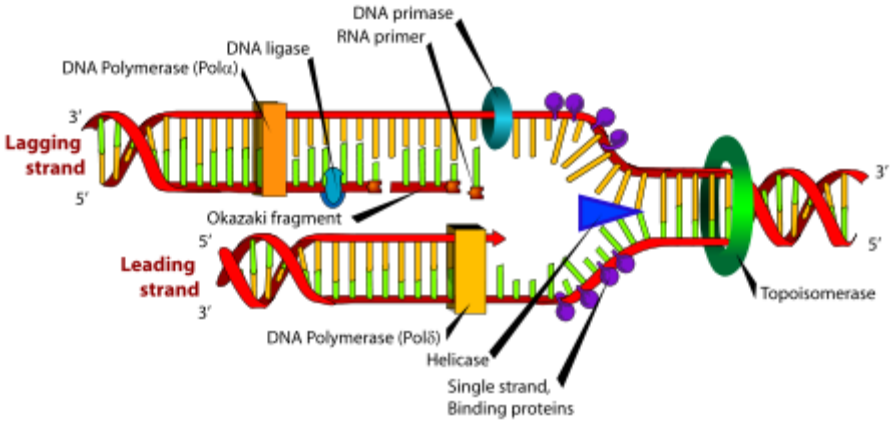
Some non-coding DNA sequences play structural roles in chromosomes. Telomeres and centromeres typically contain few genes, but are important for the function and stability of chromosomes.[43][74] An abundant form of non-coding DNA in humans are pseudogenes, which are copies of genes that have been disabled by mutation.[75] 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.[76]

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.



DNA replication. The double helix is unwound by a helicase and topoisomerase. Next, one DNA polymerase produces the leading strand copy. Another DNA polymerase binds to the lagging strand. This enzyme makes discontinuous segments (called Okazaki fragments) before DNA ligase joins them together.

Replication

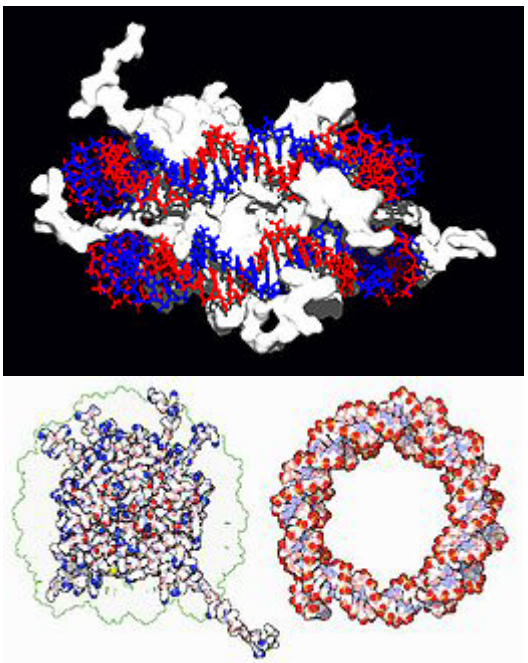
Cell division is essential for an organism to grow, but when a cell divides it must replicate the DNA in its genome so that the two daughter cells have the same genetic information as their parent. The double-stranded structure of DNA provides a simple mechanism for DNA replication. Here, the two strands 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.[77] 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.

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.

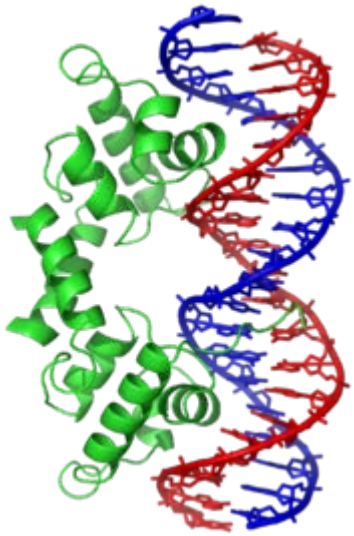
DNA-binding proteins



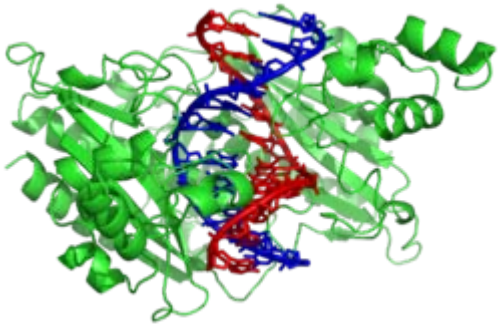
Interaction of DNA with histones (shown in white, top). These proteins' basic amino acids (below left, blue) bind to the acidic phosphate groups on DNA (below right, red).

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.[78][79] 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.[80] Chemical modifications of these basic amino acid residues include methylation, phosphorylation and acetylation.[81] 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.[82] Other non-specific DNA-binding proteins in chromatin include the high-mobility group proteins, which bind to bent or distorted DNA.[83] These proteins are important in bending arrays of nucleosomes and arranging them into the larger structures that make up chromosomes.[84]

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.[85] These binding proteins seem to stabilize single-stranded DNA and protect it from forming stem-loops or being degraded by nucleases.



The lambda repressor helix-turn-helix transcription factor bound to its DNA target[86]
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.[87] Alternatively, transcription factors can bind enzymes that modify the histones at the promoter; this will change the accessibility of the DNA template to the polymerase.[88]
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.[89] 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.[90]



The restriction enzyme EcoRV (green) in a complex with its substrate DNA[91]

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′-GAT|ATC-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.[92] 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.[93] 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.[93]

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.[27] 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.[94] Topoisomerases are required for many processes involving DNA, such as DNA replication and transcription.[28]

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.[95] 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 copies of existing polynucleotide chains - which are called templates. These enzymes function by adding nucleotides onto the 3′ hydroxyl group of the previous nucleotide in a DNA strand. Consequently, all polymerases work in a 5′ to 3′ direction.[96] 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.

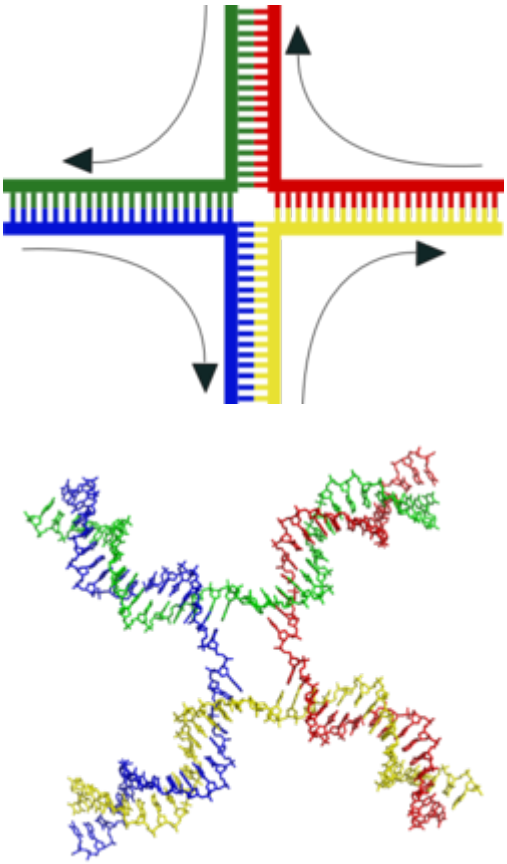
In DNA replication, a DNA-dependent DNA polymerase makes a copy of a DNA sequence. Accuracy is vital in this process, so many of these polymerases have a proofreading activity. Here, the polymerase recognizes the occasional mistakes in the synthesis reaction by the lack of base pairing between the mismatched nucleotides. If a mismatch is detected, a 3′ to 5′ exonuclease activity is activated and the incorrect base removed.[97] 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.[98]

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.[42][99] Telomerase is an unusual polymerase because it contains its own RNA template as part of its structure.[43]

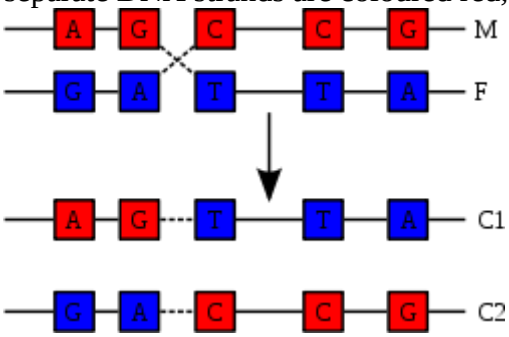
Transcription is carried out by a DNA-dependent RNA polymerase that copies the sequence of a DNA strand into RNA. To begin transcribing a gene, the RNA polymerase binds to a sequence of DNA called a promoter and separates the DNA strands. It then copies the gene sequence into a messenger RNA transcript 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.[100]

Genetic recombination



Structure of the Holliday junction intermediate in genetic recombination. The four separate DNA strands are coloured red, blue, green and yellow.[101]



Recombination involves the breakage and rejoining of two chromosomes (M and F) to produce two re-arranged chromosomes (C1 and C2).

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".[102] 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 during chromosomal crossover when they recombine. 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.[103] Genetic recombination can also be involved in DNA repair, particularly in the cell's response to double-strand breaks.[104] 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.[105] The first step in recombination is a double-stranded break either caused by an endonuclease or damage to the DNA.[106] 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.[107]

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.[96][108] 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.[109] 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.[110]

Unfortunately, there is no direct evidence of ancient genetic systems, as recovery of DNA from most fossils is impossible. This is because DNA will survive in the environment for less than one million years and slowly degrades into short fragments in solution.[111] 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,[112] but these claims are controversial.[113][114]

Uses in technology

Genetic engineering

Methods have been developed to purify DNA from organisms, such as phenol-chloroform extraction and 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.[115] The genetically modified organisms produced can be used to produce products such as recombinant proteins, used in medical research,[116] or be grown in agriculture.[117][118]

Forensics

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 called genetic fingerprinting, or more accurately, DNA profiling. In DNA profiling, the lengths of variable sections of repetitive DNA, such as short tandem repeats and

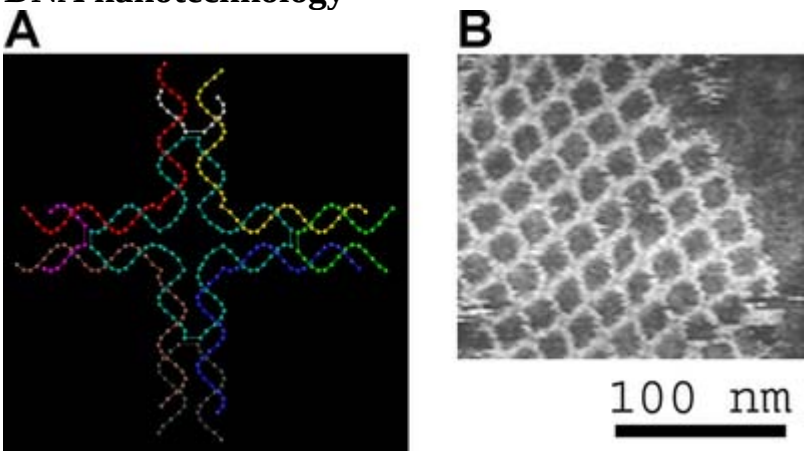
minisatellites, are compared between people. This method is usually an extremely reliable technique for identifying a matching DNA.[119] However, identification can be complicated if the scene is contaminated with DNA from several people.[120] DNA profiling was developed in 1984 by British geneticist Sir Alec Jeffreys,[121] and first used in forensic science to convict Colin Pitchfork in the 1988 Enderby murders case.[122]

People convicted of certain types of crimes may be required to provide a sample of DNA for a database. This has helped investigators solve old cases where only a DNA sample was obtained from the scene. DNA profiling can also be used to identify victims of mass casualty incidents.[123] On the other hand, many convicted people have been released from prison on the basis of DNA techniques, which were not available when a crime had originally been committed.

Bioinformatics

Bioinformatics involves the manipulation, searching, and data mining of DNA sequence data. The development of techniques to store and search DNA sequences have led to widely applied advances in computer science, especially string searching algorithms, machine learning and database theory.[124] 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.[125] In other applications such as text editors, even simple algorithms for this problem usually suffice, but DNA sequences cause these algorithms to exhibit near-worst-case behaviour due to their small number of distinct characters. The related problem of sequence alignment aims 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.[126] Data sets representing entire genomes' worth of DNA sequences, such as those produced by the Human Genome Project, are difficult to use without annotations, which label 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 in an organism even before they have been isolated experimentally.[127]

DNA nanotechnology



The DNA structure at left (schematic shown) will self-assemble into the structure visualized by atomic force microscopy at right. DNA nanotechnology is the field which seeks to design nanoscale structures using the molecular recognition properties of DNA molecules. Image from Strong, 2004.

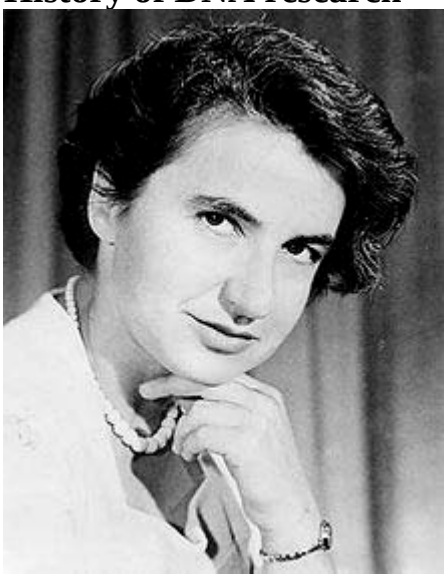
DNA nanotechnology uses the unique molecular recognition properties of DNA and other nucleic acids to create self-assembling branched DNA complexes with useful properties.[128] DNA is thus used as a structural material rather than as a carrier of biological information. This has led to the creation of two-dimensional periodic lattices (both tile-based as well as using the "DNA origami" method) as well as three-dimensional structures in the shapes of polyhedra.[129] Nanomechanical devices and algorithmic self-assembly have also been demonstrated,[130] and these DNA structures have been used to template the arrangement of other molecules such as gold nanoparticles and streptavidin proteins.[131]

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.[132] 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.[133][134]

DNA has also been used to look at modern family relationships, such as establishing family relationships between the descendants of Sally Hemings and Thomas Jefferson. This usage is closely related to the use of DNA in criminal investigations detailed above. Indeed, some criminal investigations have been solved when DNA from crime scenes has matched relatives of the guilty individual.[135]

History of DNA research



Rosalind Franklin, co-creator of the single X-ray diffraction image

Raymond Gosling, co-creator of the single X-ray diffraction image

Francis Crick, co-originator of the double-helix model

Further information: History of molecular biology

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".[136] In 1919, Phoebus Levene identified the base, sugar and phosphate nucleotide unit.[137] Levene suggested that DNA consisted of a string of nucleotide units linked together through the phosphate groups. However, 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.[138]

In 1928, Frederick Griffith discovered that traits of the "smooth" form of the Pneumococcus could be transferred to the "rough" form of the same bacteria by mixing killed "smooth" bacteria with the live "rough" form.[139] This system provided the first clear suggestion that DNA carried 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.[140] 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.[141]

In 1953 James D. Watson and Francis Crick suggested what is now accepted as the first correct double-helix model of DNA structure in the journal Nature.[7] Their double-helix, molecular model of DNA was then based on a single X-ray diffraction image (labeled as "Photo 51")[142] taken by Rosalind Franklin and Raymond Gosling in May 1952, as well as the information that the DNA bases were paired—also obtained through private

communications from Erwin Chargaff in the previous years. Chargaff's rules played a very important role in establishing double-helix configurations for B-DNA as well as A-DNA.
Experimental evidence supporting the Watson and Crick model were published in a series of five articles in the same issue of Nature.[143] 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[31][144]; 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.[32] In 1962, after Franklin's death, Watson, Crick, and Wilkins jointly received the Nobel Prize in Physiology or Medicine.[145] Unfortunately, Nobel rules of the time allowed only living recipients, but a vigorous debate continues on who should receive credit for the discovery.[146]
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".[147] Final confirmation of the replication mechanism that was implied by the double-helical structure followed in 1958 through the Meselson-Stahl experiment.[148] 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.[149] These findings represent the birth of molecular biology.

Real-time polymerase chain reaction

From Wikipedia, the free encyclopedia

For reverse transcription polymerase chain reaction (RT-PCR), see reverse transcription polymerase chain reaction.

In molecular biology, **real-time polymerase chain reaction**, also called *quantitative real time polymerase chain reaction* (Q-PCR/qPCR/qrt-PCR) or *kinetic polymerase chain reaction* (KPCR), is a laboratory technique based on the PCR, which is used to amplify and simultaneously quantify a targeted DNA molecule. It enables both detection and quantification (as absolute number of copies or relative amount when normalized to DNA input or additional normalizing genes) of one or more specific sequences in a DNA sample.

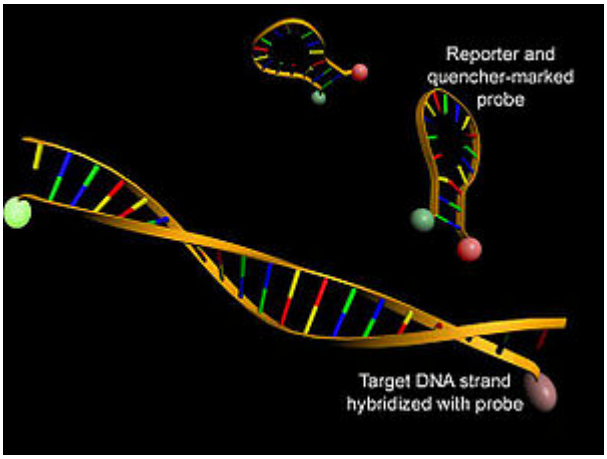
The procedure follows the general principle of polymerase chain reaction; its key feature is that the amplified DNA is detected as the reaction progresses in *real time*, a new approach compared to standard PCR, where the product of the reaction is detected at its end. Two common methods for detection of products in real-time PCR are: (1) non-specific fluorescent dyes that intercalate with any double-stranded DNA, and (2) sequence-specific DNA probes consisting of oligonucleotides that are labeled with a fluorescent reporter which permits detection only after hybridization of the probe with its complementary DNA target.

Frequently, real-time PCR is combined with reverse transcription to quantify messenger RNA and Non-coding RNA in cells or tissues.

Abbreviations used for real-time PCR methods vary widely and include RTQ-PCR, Q-PCR or qPCR. [1] Real-time reverse-transcription PCR is often denoted as qRT-PCR,[2], RRT-PCR,[3] or RT-rt PCR.[4] The acronym RT-PCR commonly denotes reverse-transcription PCR and not real-time PCR, but not all authors adhere to this convention.[5]

Contents

Background



Real time quantitative PCR uses fluorophores in order to detect levels of gene expression.

Cells in all organisms regulate gene expression and turnover of gene transcripts (messenger RNA, abbreviated to mRNA), and the number of copies of an mRNA transcript of a gene in a cell or tissue is determined by the rates of its expression and degradation.

Northern blotting is often used to estimate the expression level of a gene by visualizing the abundance of its mRNA transcript in a sample. In this method, purified RNA is separated by agarose gel electrophoresis, transferred to a solid matrix (such as a nylon membrane), and probed with a specific DNA or RNA probe that is complementary to the gene of interest. Although this technique is still used to assess gene expression, it requires relatively large amounts of RNA and provides only qualitative or semiquantitative information of mRNA levels.

In order to robustly detect and quantify gene expression from small amounts of RNA, amplification of the gene transcript is necessary. The polymerase chain reaction is a common method for amplifying DNA; for mRNA-based PCR the RNA sample is first reverse transcribed to cDNA with reverse transcriptase.

Development of PCR technologies based on reverse transcription and *fluorophores* permits measurement of DNA amplification during PCR in real time, i.e., the amplified product is measured at each PCR cycle. The data thus generated can be analysed by computer software to calculate *relative gene expression* in several samples, or *mRNA copy number*. Real-time PCR can also be applied to the detection and quantification of DNA in samples to determine the presence and abundance of a particular DNA sequence in these samples.

Real-time PCR with double-stranded DNA-binding dyes as reporters

A DNA-binding dye binds to all double-stranded (ds)DNA in PCR, causing fluorescence of the dye. An increase in DNA product during PCR therefore leads to an increase in fluorescence intensity and is measured at each cycle, thus allowing DNA concentrations to be quantified. However, dsDNA dyes such as SYBR Green will bind to all dsDNA PCR products, including nonspecific PCR products (such as Primer dimer). This can potentially interfere with or prevent accurate quantification of the intended target sequence.

1. The reaction is prepared as usual, with the addition of fluorescent dsDNA dye.
2. The reaction is run in a Real-time PCR instrument, and after each cycle, the levels of fluorescence are measured with a detector; the dye only fluoresces when bound to the dsDNA (i.e., the PCR product). With reference to a standard dilution, the dsDNA concentration in the PCR can be determined.

Like other real-time PCR methods, the values obtained do not have absolute units associated with them (i.e., mRNA copies/cell). As described above, a comparison of a measured DNA/RNA sample to a standard dilution will only give a fraction or ratio of the sample relative to the standard, allowing only relative comparisons between different tissues or experimental conditions. To ensure accuracy in the quantification, it is usually necessary to normalize expression of a target gene to a stably expressed gene (see below).

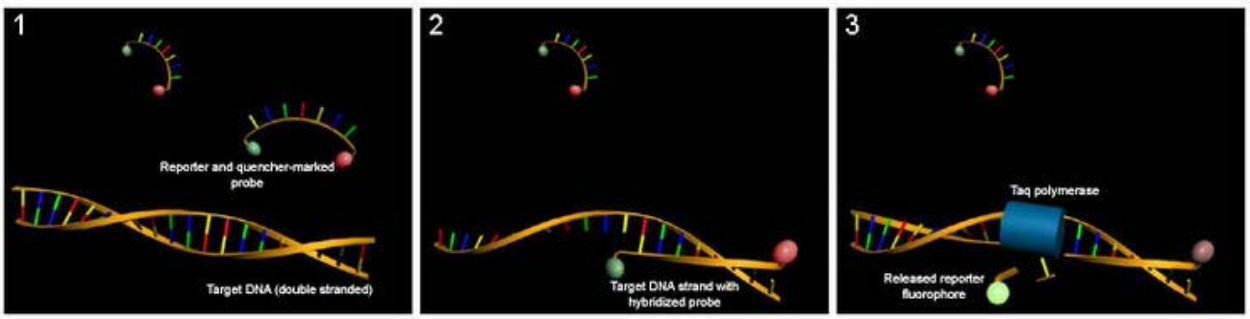
This can correct possible differences in RNA quantity or quality across experimental samples.

Fluorescent reporter probe method

Fluorescent reporter probes detect only the DNA containing the probe sequence; therefore, use of the reporter probe significantly increases specificity, and enables quantification even in the presence of non-specific DNA amplification. Fluorescent probes can be used in multiplex assays—for detection of several genes in the same reaction—based on specific probes with different-coloured labels, provided that all targeted genes are amplified with similar efficiency. The specificity of fluorescent reporter probes also prevents interference of measurements caused by primer dimers, which are undesirable potential by-products in PCR. However, fluorescent reporter probes do not prevent the inhibitory effect of the primer dimers, which may depress accumulation of the desired products in the reaction.

The method relies on a DNA-based probe with a fluorescent reporter at one end and a quencher of fluorescence at the opposite end of the probe. The close proximity of the reporter to the quencher prevents detection of its fluorescence; breakdown of the probe by the 5' to 3' exonuclease activity of the Taq polymerase breaks the reporter-quencher proximity and thus allows unquenched emission of fluorescence, which can be detected after excitation with a laser. An increase in the product targeted by the reporter probe at each PCR cycle therefore causes a proportional increase in fluorescence due to the breakdown of the probe and release of the reporter.

1. The PCR is prepared as usual (see PCR), and the reporter probe is added.
2. As the reaction commences, during the annealing stage of the PCR both probe and primers anneal to the DNA target.
3. Polymerisation of a new DNA strand is initiated from the primers, and once the polymerase reaches the probe, its 5'-3'-exonuclease degrades the probe, physically separating the fluorescent reporter from the quencher, resulting in an increase in fluorescence.
4. Fluorescence is detected and measured in the real-time PCR thermocycler, and its geometric increase corresponding to exponential increase of the product is used to determine the threshold cycle (C_T) in each reaction.



(1) In intact probes, reporter fluorescence is quenched. (2) Probes and the complementary DNA strand are hybridized and reporter fluorescence is still quenched. (3) During PCR, the probe is degraded by the Taq polymerase and the fluorescent reporter released.

Quantification

Quantifying gene expression by traditional methods presents several problems. Firstly, detection of mRNA on a Northern blot or PCR products on a gel or Southern blot is time-consuming and does not allow precise quantification. Also, over the 20-40 cycles of a typical PCR, the amount of product reaches a plateau determined more by the amount of primers in the reaction mix than by the input template/sample.

Relative concentrations of DNA present during the exponential phase of the reaction are determined by plotting fluorescence against cycle number on a logarithmic scale (so an exponentially increasing quantity will give a straight line). A threshold for detection of fluorescence above background is determined. The cycle at which the fluorescence from a sample crosses the threshold is called the cycle threshold, C_t. The quantity of DNA theoretically doubles every cycle during the exponential phase and relative amounts of DNA can be calculated, e.g. a sample whose C_t is 3 cycles earlier than another's has $2^3 = 8$ times more template. Since all sets of primers don't work equally well, one has to calculate the reaction efficiency first. Thus, by using this as the base and the cycle difference $C(_t)$ as the exponent, the precise difference in starting template can be calculated (in previous example, if efficiency was 1.96, then the sample would have 7.53 times more template).

Amounts of RNA or DNA are then determined by comparing the results to a standard curve produced by real-time PCR of serial dilutions (e.g. undiluted, 1:4, 1:16, 1:64) of a known amount of RNA or DNA. As mentioned above, to accurately quantify gene expression, the measured amount of RNA from the gene of interest is divided by the amount of RNA from a housekeeping gene measured in the same sample to normalize for possible variation in the amount and quality of RNA between different samples. This normalization permits accurate comparison of expression of the gene of interest between different samples, provided that the expression of the reference (housekeeping) gene used in the normalization is very similar across all the samples. Choosing a reference gene fulfilling this criterion is therefore of high importance, and often challenging, because only very few genes show equal levels of expression across a range of different conditions or tissues. [6][7]

Real-time PCR can be used to quantify nucleic acids by two strategies - Relative quantification and Absolute quantification. Relative quantification measures the fold-difference (2X, 3X etc.) in the target amount. Absolute quantification gives the exact number of target molecules present by comparing with known standards. The quality of Standard is important for accurate quantification. [8]

Applications of real-time polymerase chain reaction

There are numerous applications for real-time polymerase chain reaction in the laboratory. It is commonly used for both diagnostic and basic research.

Diagnostic real-time PCR is applied to rapidly detect nucleic acids that are diagnostic of, for example, infectious diseases, cancer and genetic abnormalities. The introduction of real-time PCR assays to the clinical microbiology laboratory has significantly improved

the diagnosis of infectious diseases, [9] and is deployed as a tool to detect newly emerging diseases, such as flu, in diagnostic tests.[10]

In research settings, real-time PCR is mainly used to provide quantitative measurements of gene transcription. The technology may be used in determining how the genetic expression of a particular gene changes over time, such as in the response of tissue and cell cultures to an administration of a pharmacological agent, progression of cell differentiation, or in response to changes in environmental conditions.