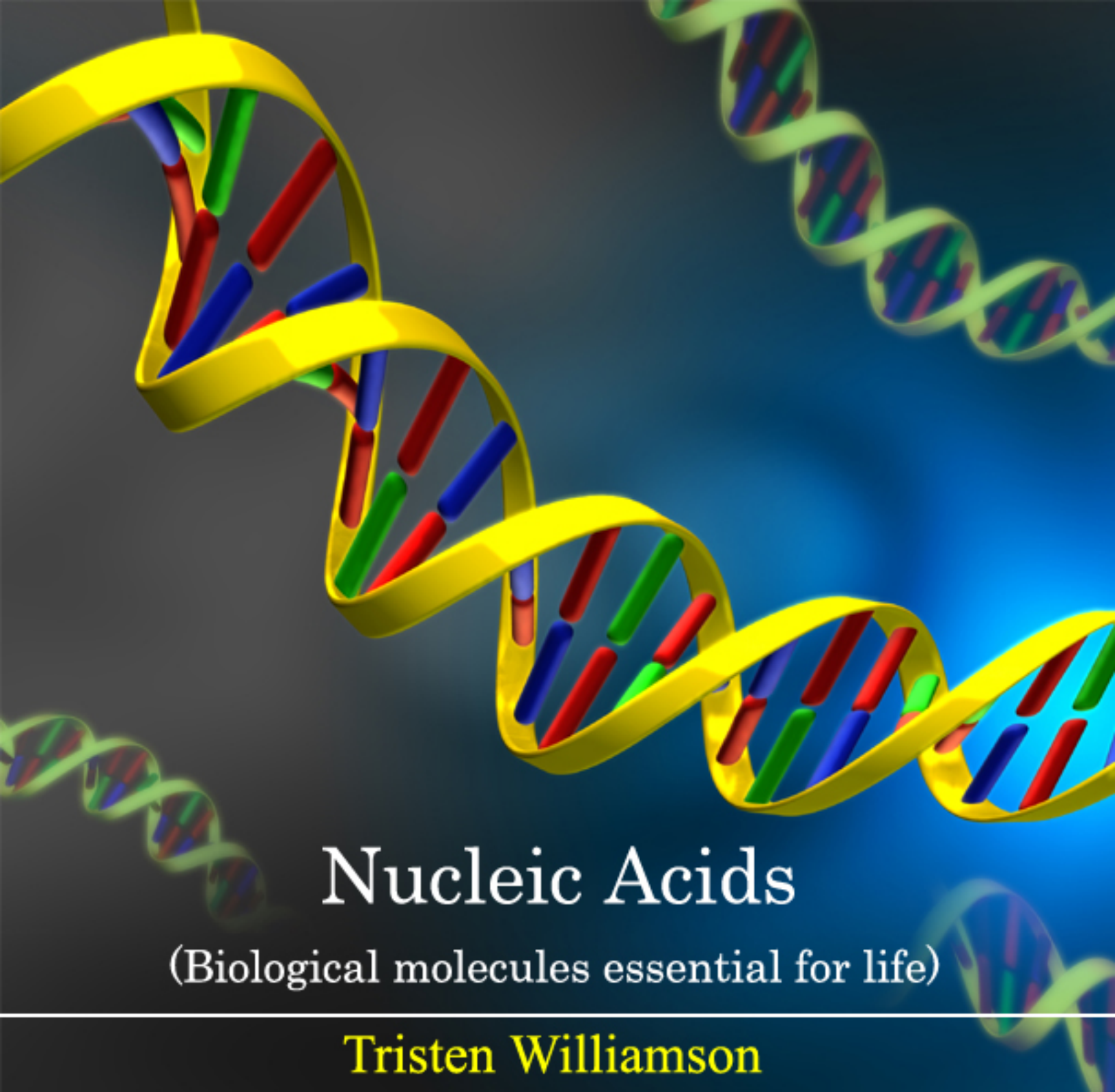


Nucleic Acids

(Biological molecules essential for life)

Tristen Williamson

First Edition, 2012

ISBN 978-81-323-3343-2

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Nucleic Acid

Nucleic acids are biological molecules essential for life, and include DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). Together with proteins, nucleic acids make up the most important macromolecules; each is found in abundance in all living things. Nucleic acids were first discovered by Friedrich Miescher in 1871. Experimental studies of nucleic acids constitute a major part of modern biological and medical research, and form a foundation for genome and forensic science, as well as the biotechnology and pharmaceutical industries.

Occurrence and nomenclature

The term *nucleic acid* is the over all name for DNA and RNA, members of a family of biopolymers, and is synonymous with *polynucleotide*. Nucleic acids were named for their initial discovery within the cell nucleus, and for the presence of phosphate groups (related to phosphoric acid). Although first discovered within the nucleus of eukaryotic cells, nucleic acids are now known to be found in all life forms, including within bacteria, archaea, mitochondria, chloroplasts, viruses and viroids. All living cells and organelles contain both DNA and RNA, while viruses contain either DNA or RNA, but not usually both. The basic component of biological nucleic acids is the nucleotide, each of which contains a pentose sugar (ribose or deoxyribose), a phosphate group, and a nucleobase. Nucleic acids are also generated within the laboratory, through the use of enzymes (DNA and RNA polymerases) and by solid-phase chemical synthesis. The chemical methods also enable the generation of altered nucleic acids that are not found in nature, for example peptide nucleic acids.

Molecular composition and size

Nucleic acids can vary in size, but are generally very large molecules. Indeed, DNA molecules are probably the largest individual molecules known. Well-studied biological nucleic acid molecules range in size from 21 nucleotides (small interfering RNA) to large chromosomes (human chromosome 1 is a single molecule that contains 247 million base pairs).

In most cases, naturally occurring DNA molecules are double-stranded and RNA molecules are single-stranded. There are numerous exceptions, however--some viruses have genomes made of double-stranded RNA and other viruses have single-stranded DNA genomes, and, in some circumstances, nucleic acid structures with three or four strands can form.

Nucleic acids are linear polymers (chains) of nucleotides. Each nucleotide consists of three components: a purine or pyrimidine nucleobase (sometimes termed *nitrogenous base* or simply *base*), a pentose sugar; and a phosphate group. The substructure consisting of a nucleobase plus sugar is termed a nucleoside. Nucleic acid types differ in the structure of the sugar in their nucleotides - DNA contains 2'-deoxyribose while RNA contains ribose (where the only difference is the presence of a hydroxyl group). Also, the nucleobases found in the two nucleic acid types are different: adenine, cytosine, and guanine are found in both RNA and DNA, while thymine occurs in DNA and uracil occurs in RNA.

The sugars and phosphates in nucleic acids are connected to each other in an alternating chain (sugar-phosphate backbone) through phosphodiester linkages. In conventional nomenclature, the carbons to which the phosphate groups attach are the 3'-end and the 5'-end carbons of the sugar. This gives nucleic acids directionality, and the ends of nucleic acid molecules are referred to as 5'-end and 3'-end. The nucleobases are joined to the sugars via an N-glycosidic linkage involving a nucleobase ring nitrogen (N-1 for pyrimidines and N-9 for purines) and the 1' carbon of the pentose sugar ring.

Non-standard nucleosides are also found in both RNA and DNA and usually arise from modification of the standard nucleosides within the DNA molecule or the primary (initial) RNA transcript. Transfer RNA (tRNA) molecules contain a particularly large number of modified nucleosides.

Topology

Double-stranded nucleic acids are made up of complementary sequences, in which extensive Watson-Crick base pairing results in the formation of a highly repeated and quite uniform double-helical three-dimensional structure. In contrast, single-stranded RNA and DNA molecules are not constrained to a regular double helix, and can adopt highly complex three-dimensional structures that are based on short stretches of intramolecular base-paired sequences that include both Watson-Crick and noncanonical base pairs, as well as a wide range of complex tertiary interactions.

Nucleic acid molecules are usually unbranched, and may occur as linear and circular molecules. For example, bacterial chromosomes, plasmids, mitochondrial DNA and chloroplast DNA are usually circular double-stranded DNA molecules, while chromosomes of the eukaryotic nucleus are usually linear double-stranded DNA molecules. Most RNA molecules are linear, single-stranded molecules, but both circular and branched molecules can result from RNA splicing reactions.

Nucleic acid sequences

One DNA or RNA molecule differs from another primarily in the sequence of nucleotides. Nucleotide sequences are of great importance in biology, since they carry the ultimate instructions that encode all biological molecules, molecular assemblies, subcellular and cellular structures, organs and organisms, and directly enable cognition, memory and behavior (See: Genetics). Enormous efforts have gone into the development of experimental methods to determine the nucleotide sequence of biological DNA and RNA molecules, and today hundreds of millions of nucleotides are sequenced daily at genome centers and smaller laboratories worldwide.

Types of nucleic acids

Deoxyribonucleic acid

Deoxyribonucleic acid is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. The main role of DNA molecules is the long-term storage of information and DNA is often compared to a set of blueprints, 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.

Ribonucleic acid

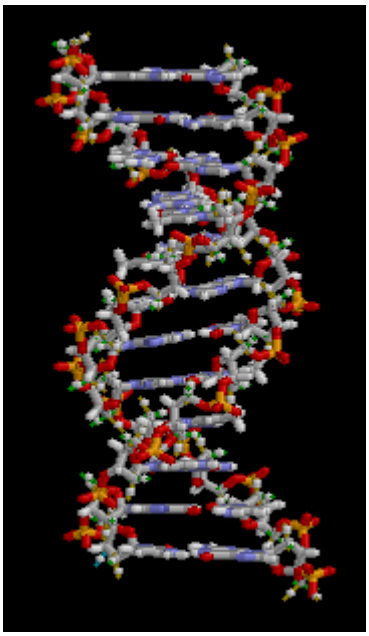
Ribonucleic acid (RNA) functions in converting genetic information from genes into the amino acid sequences of proteins. The three universal types of RNA include transfer RNA (tRNA), messenger RNA (mRNA), and ribosomal RNA (rRNA). Messenger RNA acts to carry genetic sequence information between DNA and ribosomes, directing protein synthesis. Ribosomal RNA is a major component of the ribosome, and catalyzes peptide bond formation. Transfer RNA serves as the carrier molecule for amino acids to be used in protein synthesis, and is responsible for decoding the mRNA. In addition, many other classes of RNA are now known.

Artificial nucleic acid analogs

Artificial nucleic acid analogs have been designed and synthesized by chemists, and include peptide nucleic acid, morpholino- and locked nucleic acid, as well as glycol nucleic acid and threose nucleic acid. Each of these is distinguished from naturally-occurring DNA or RNA by changes to the backbone of the molecule.

Chapter- 2

DNA



The structure of part of a DNA double helix

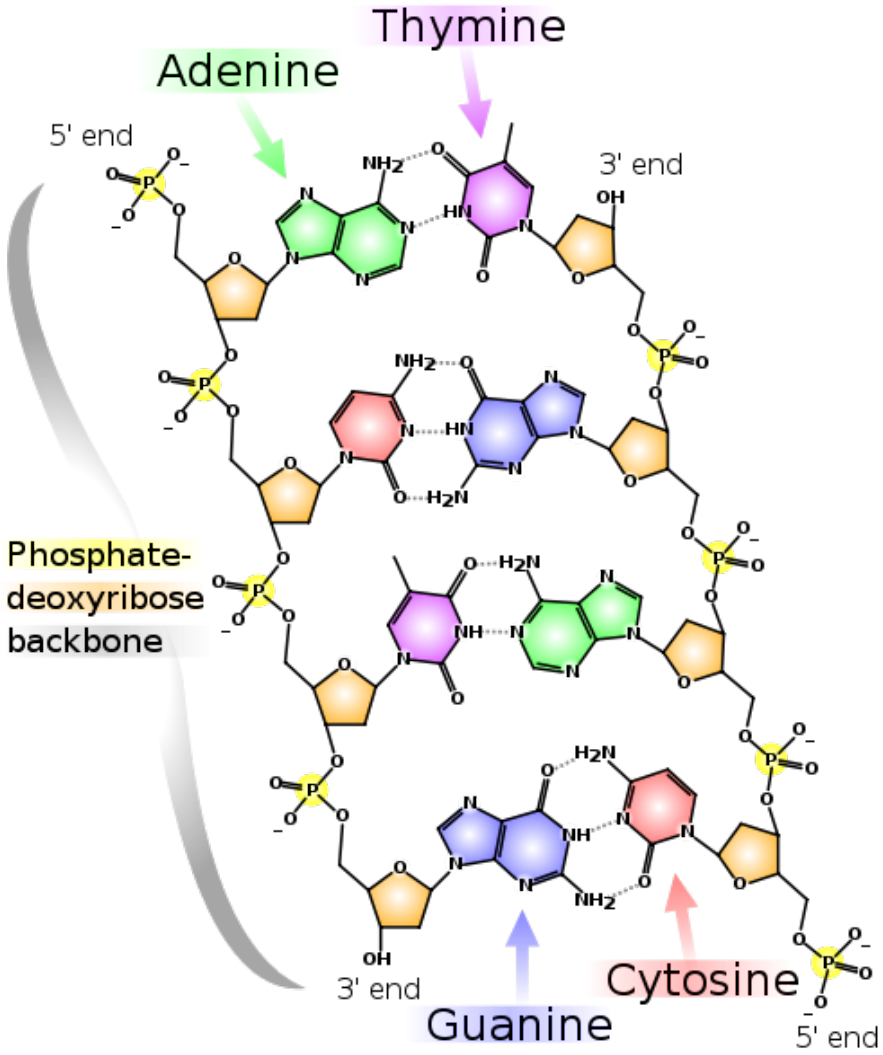
Deoxyribonucleic acid or **DNA**, is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms (with the exception of RNA viruses). The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set 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.

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. 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.

Properties

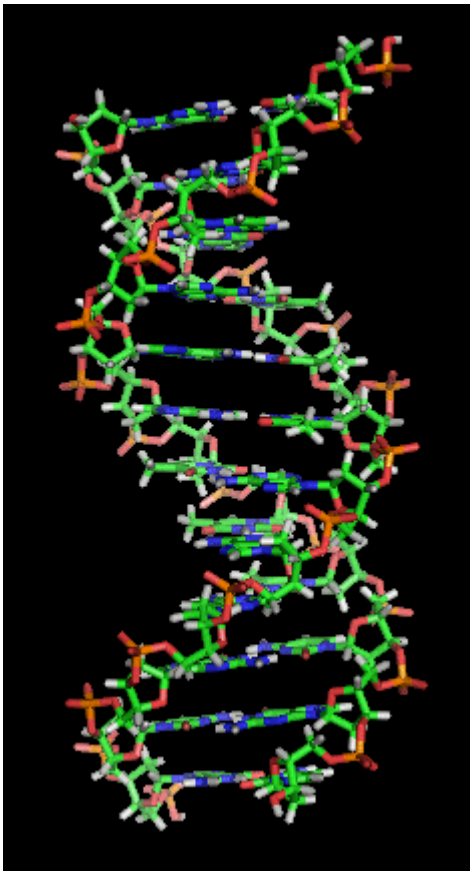


Chemical structure of DNA. Hydrogen bonds shown as dotted lines

DNA is a long polymer made from repeating units called nucleotides. As first discovered by James D. Watson and Francis Crick, the structure of DNA of 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 nanometres). 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 largest human chromosome, chromosome number 1, is approximately 220 million base pairs long.

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 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.

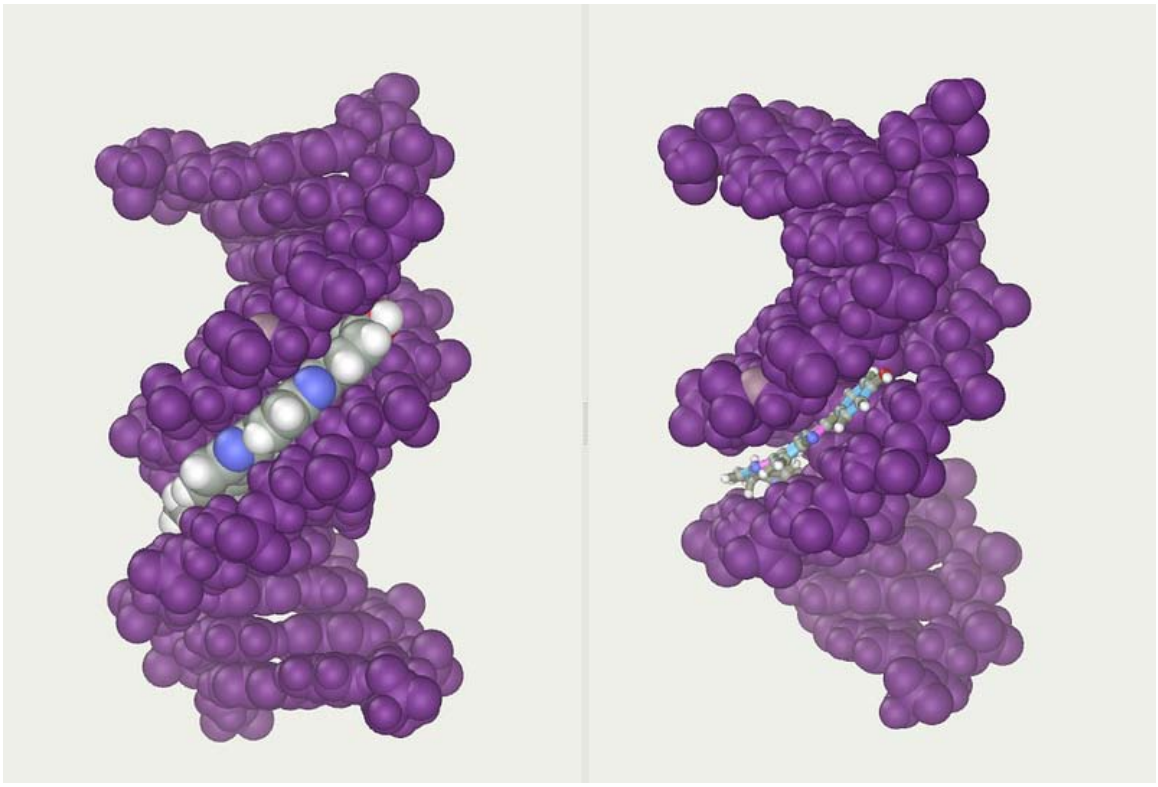
The backbone of the DNA strand is made from alternating phosphate and sugar residues. The sugar in DNA is 2-deoxyribose, which is a pentose (five-carbon) sugar. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar 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.



A section of DNA. The bases lie horizontally between the two spiraling strands.

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. 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.



Major and minor grooves of DNA. Minor groove is a binding site for the dye Hoechst 33258.

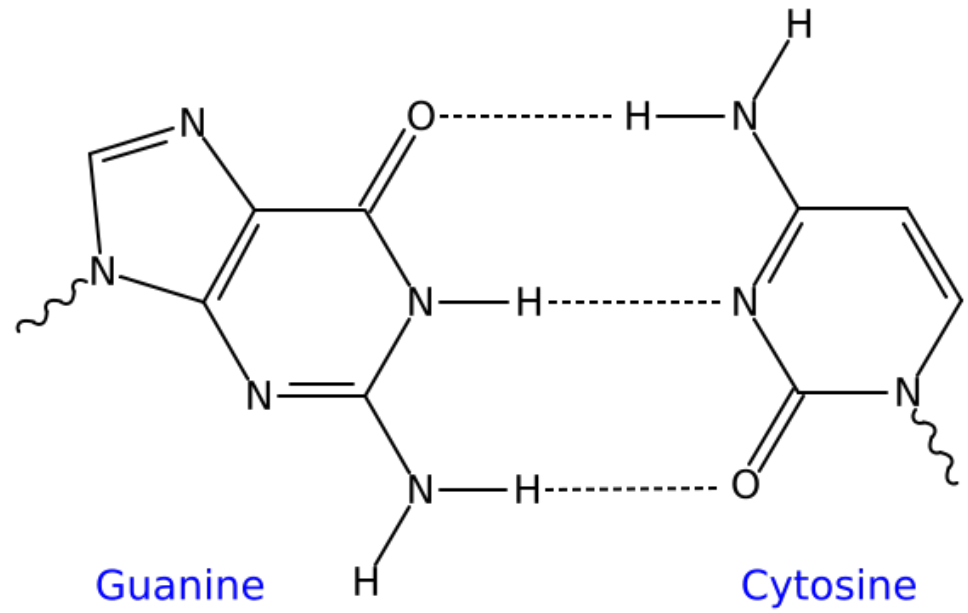
Grooves

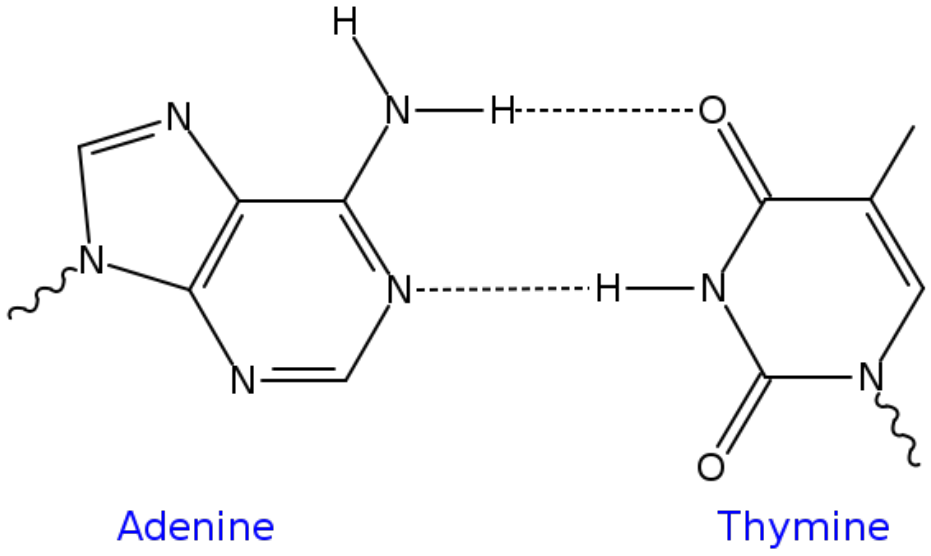
Twin helical strands form the DNA backbone. Another double helix may be found by tracing the spaces, or grooves, between the strands. These voids are adjacent to the base pairs and may provide a binding site. As the strands are not directly opposite each other, the grooves are unequally sized. One groove, the major groove, is 22 Å wide and the other, the minor groove, is 12 Å wide. 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. 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. 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.





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. 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).

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. 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 required to break the hydrogen bonds, their melting temperature (also called T_m 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

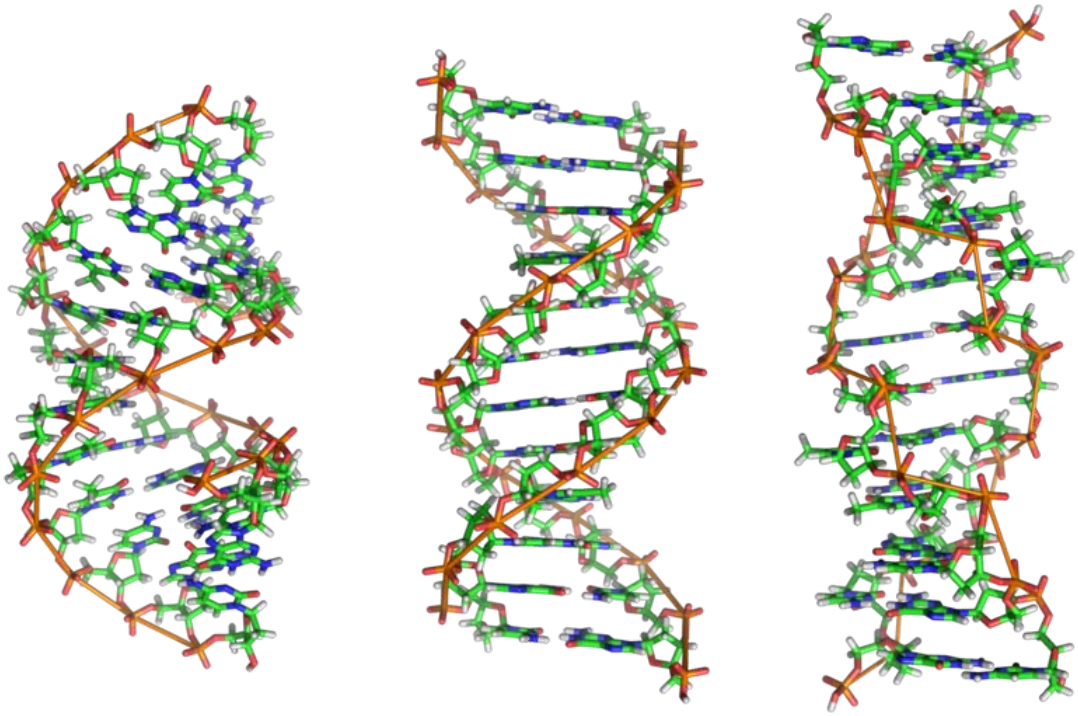
A DNA sequence is called "sense" if its sequence is the same as that of a messenger RNA copy that is translated into protein. The sequence on the opposite strand is called the "antisense" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands 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.



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. 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 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. 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.

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.

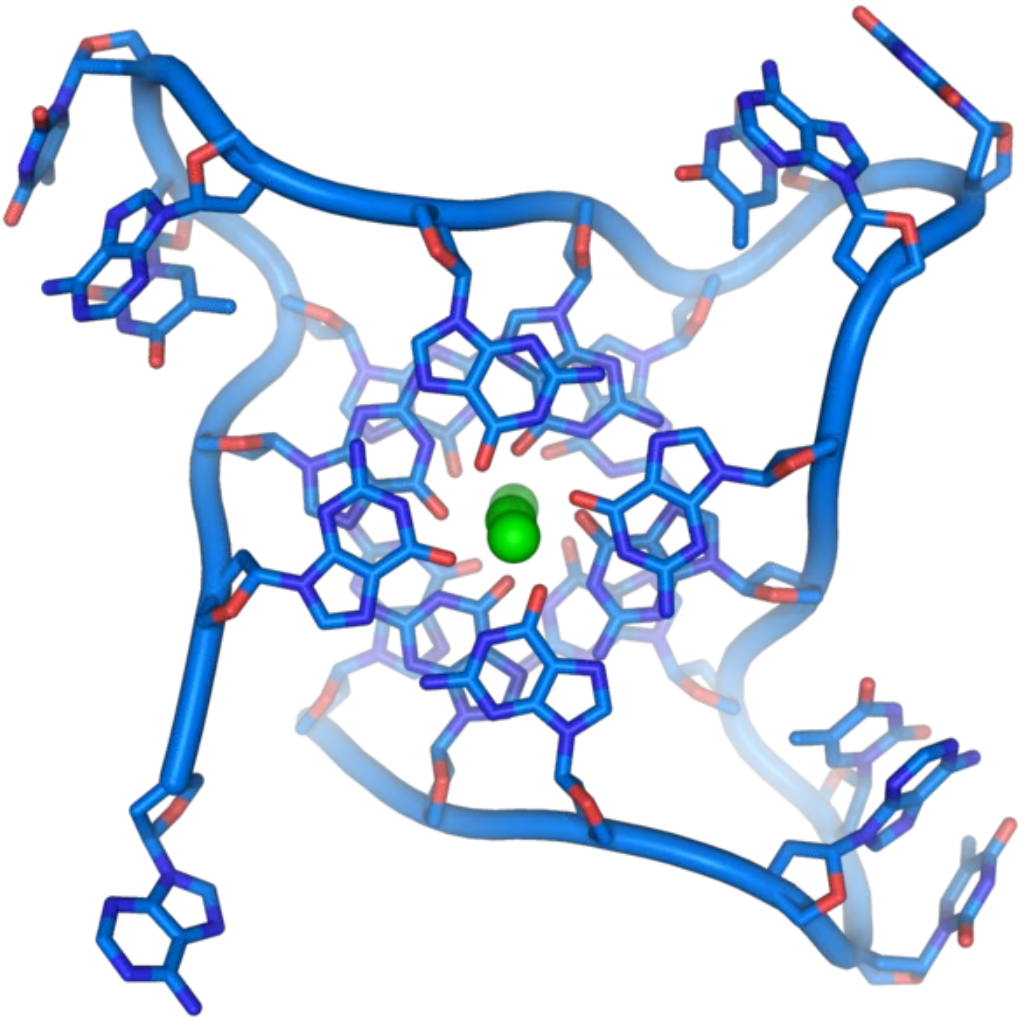
Alternate DNA chemistry

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 December 2010 NASA press conference revealed that the bacterium GFAJ-1, which has evolved in an arsenic-rich environment, is the first terrestrial lifeform found which may have this ability. The bacterium was found in Mono Lake, east of Yosemite National Park. GFAJ-1 is a rod-shaped extremophile bacterium in the family Halomonadaceae that, when starved of phosphorus, may be capable of incorporating the usually poisonous element arsenic in its DNA. This discovery lends weight to the long-standing idea that extraterrestrial life could have a different chemical makeup from life on Earth. The research was carried out by a team led by Felisa Wolfe-Simon, a geomicrobiologist and geobiochemist, a Postdoctoral Fellow of the NASA Astrobiology Institute with Arizona State University.

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.

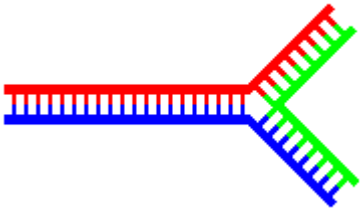


DNA quadruplex formed by telomere repeats. The looped conformation of the DNA backbone is very different from the typical DNA helix.

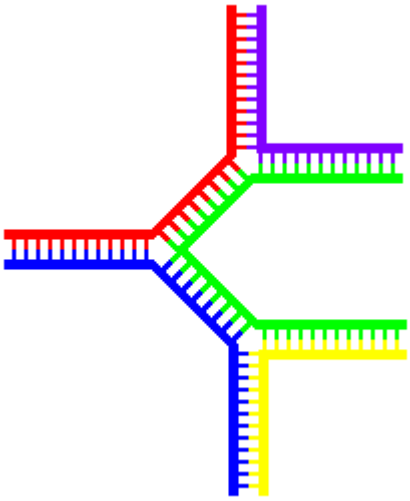
These guanine-rich sequences may stabilize chromosome ends by forming structures of stacked sets of four-base units, rather than the usual base pairs found in other DNA molecules. Here, four guanine bases form a flat plate and these flat four-base units then stack on top of each other, to form a stable *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.



Single branch



Multiple branches

Branched DNA can form networks containing multiple branches.

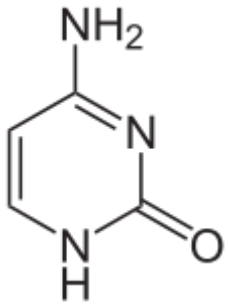
Branched DNA

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. Branched DNA can be used in nanotechnology to construct geometric shapes, see the section on uses in technology below.

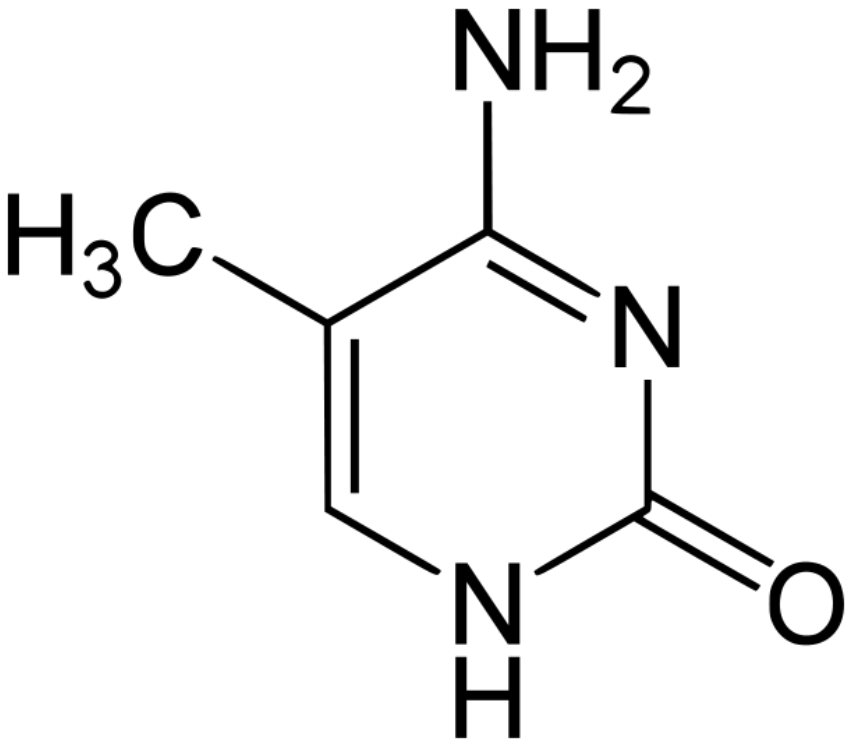
Vibration

DNA may carry out low-frequency collective motion as observed by the Raman spectroscopy and analyzed with a quasi-continuum model.

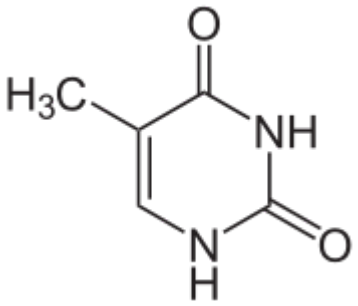
Chemical modifications



cytosine



5-methylcytosine



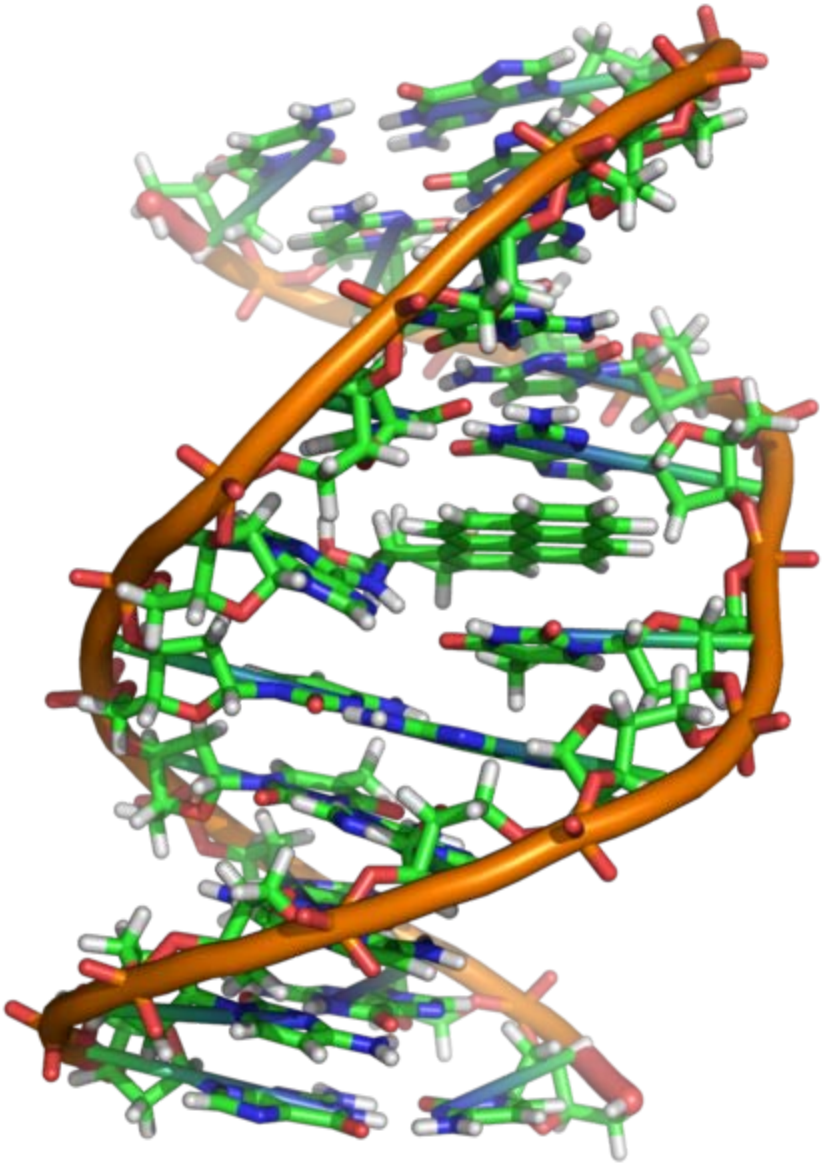
thymine

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. 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



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

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.

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. 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

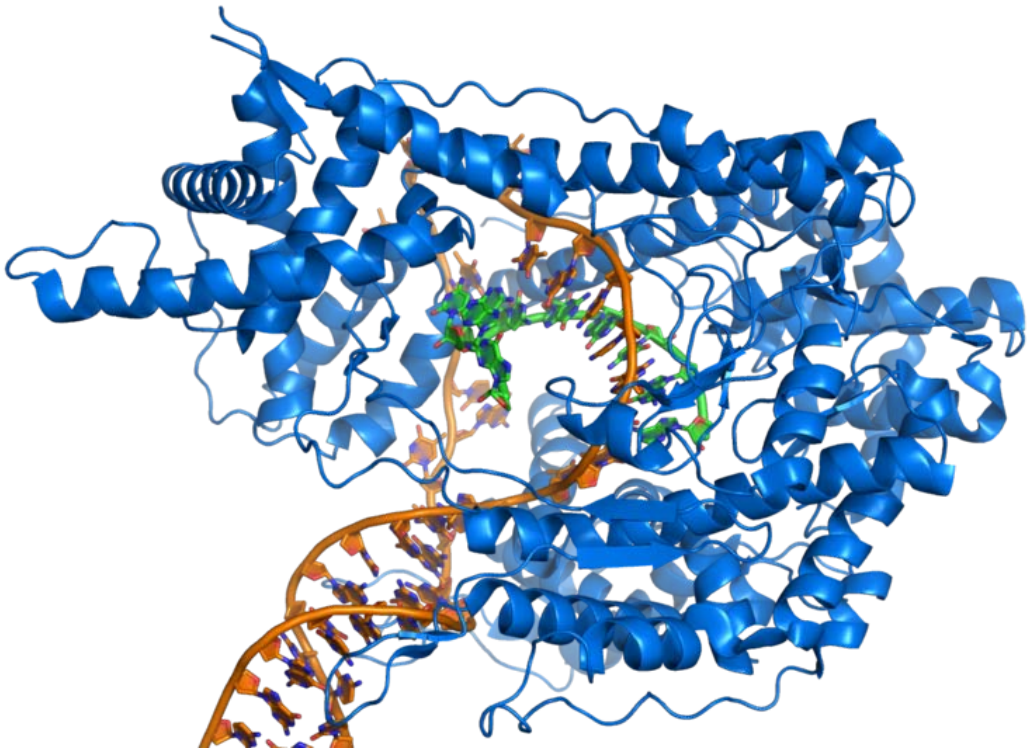
DNA usually occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The set of chromosomes in a cell makes up its genome; the human genome has approximately 3 billion base pairs of DNA arranged into 46 chromosomes. The information carried by DNA is held in the sequence of 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 we focus 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, DNA sequences that do not code protein may still encode functional non-coding RNA molecules, which are involved in the regulation of gene expression.



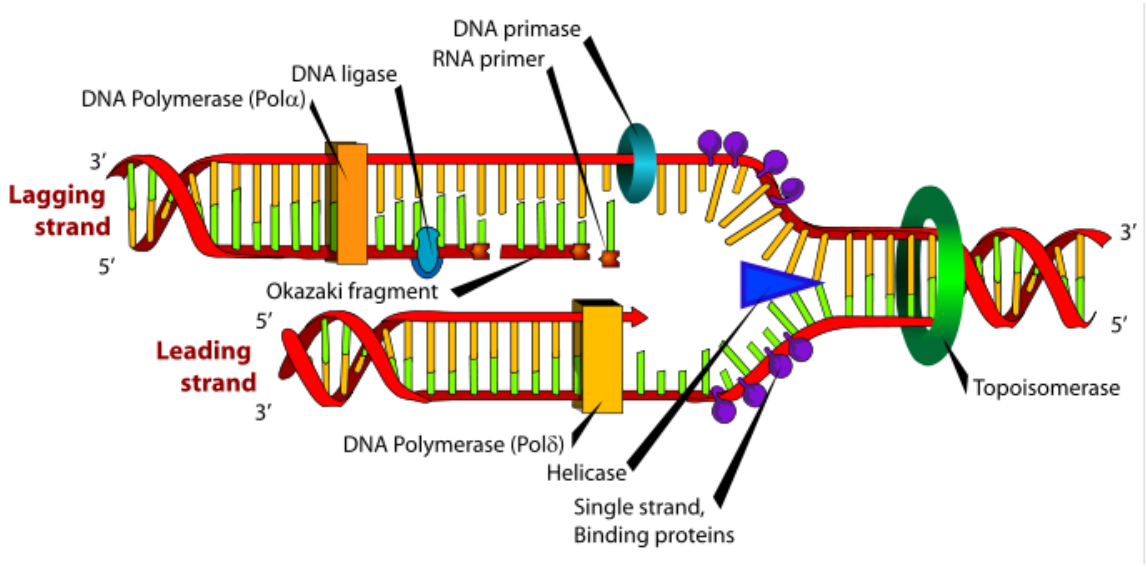
T7 RNA polymerase (blue) producing a mRNA (green) from a DNA template (orange).

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 (4^3 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.

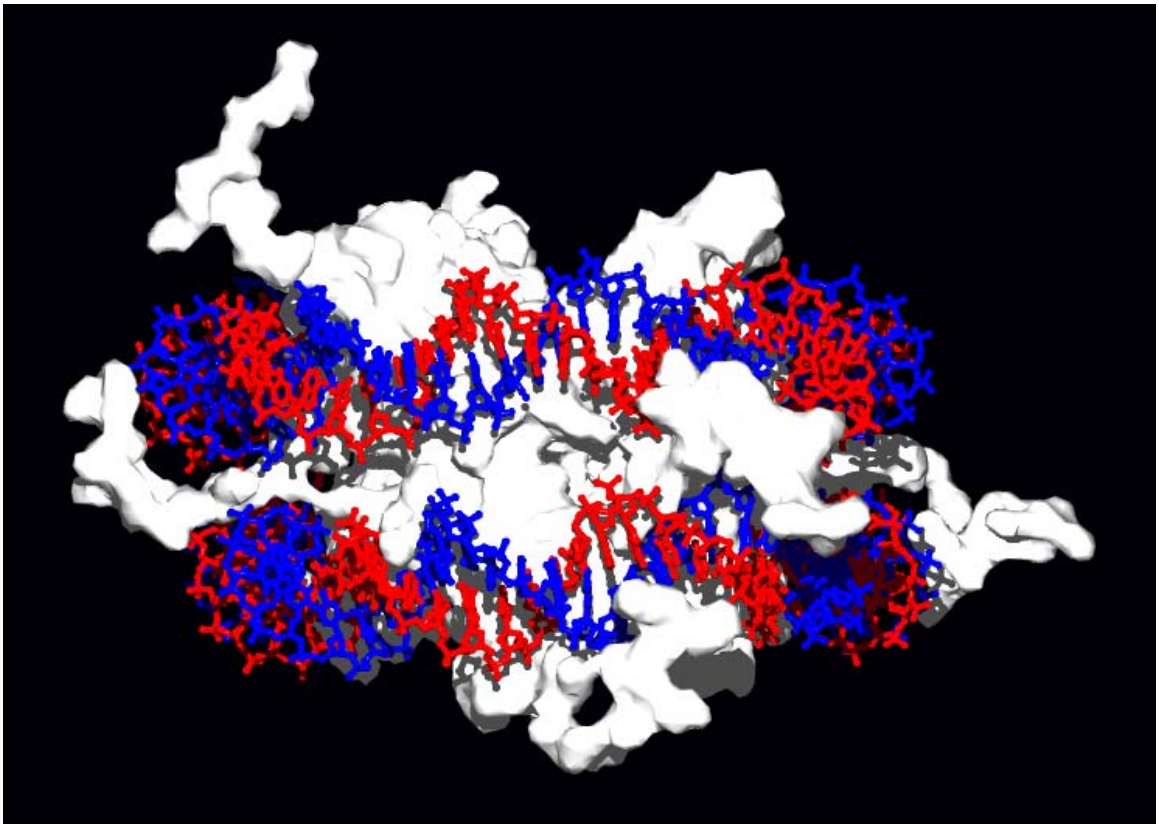
Replication

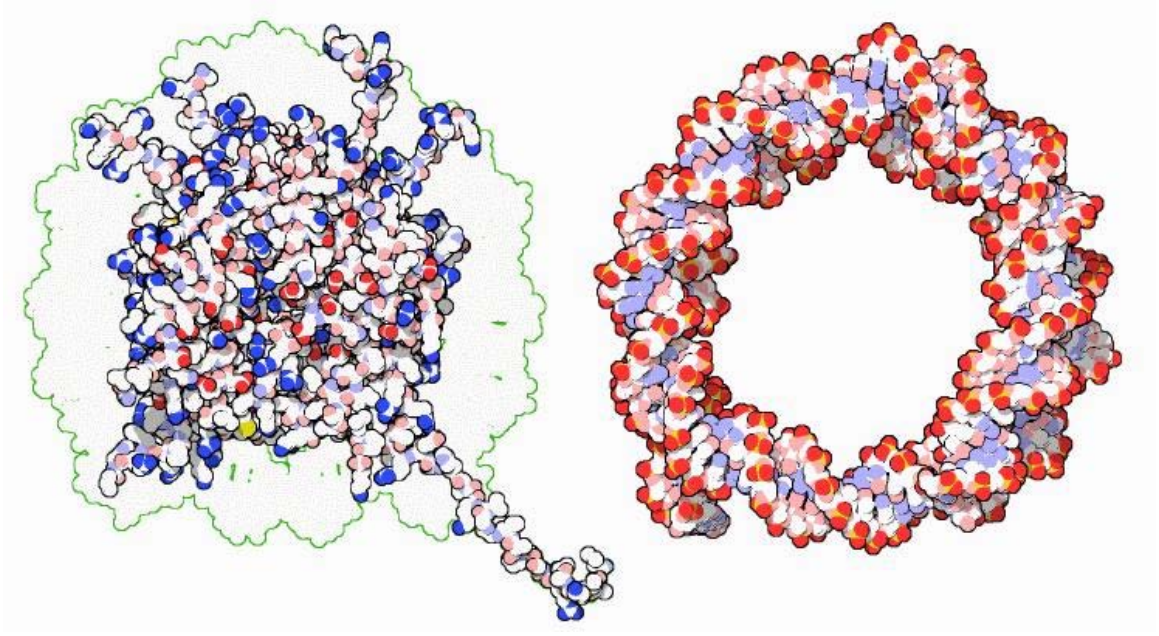
Cell division is essential for an organism to grow, but, when a cell divides, it must replicate the DNA in its genome so that the two daughter cells have the same genetic information as their parent. The double-stranded structure of DNA provides a simple mechanism for DNA replication. Here, the two strands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

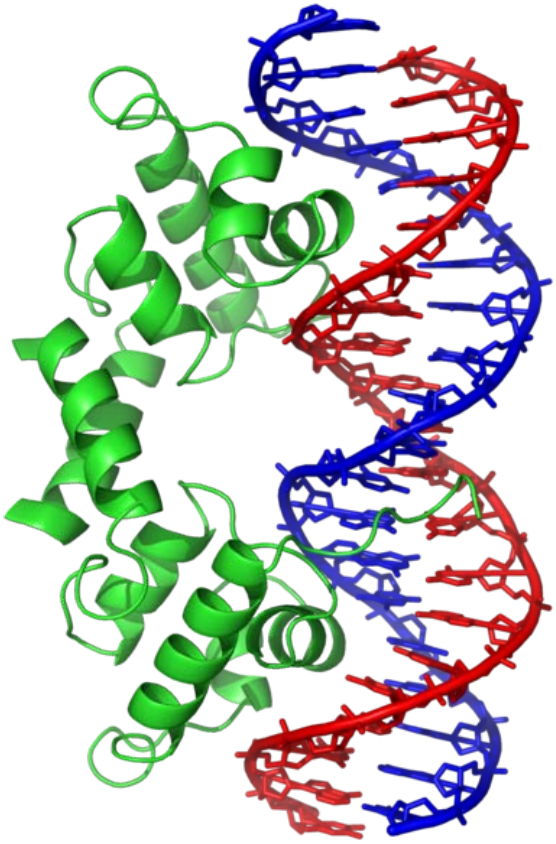




Interaction of DNA with histones (shown in white, top). These proteins' basic amino acids (below left, blue) bind to the acidic phosphate groups on DNA (below right, red).

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. 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.

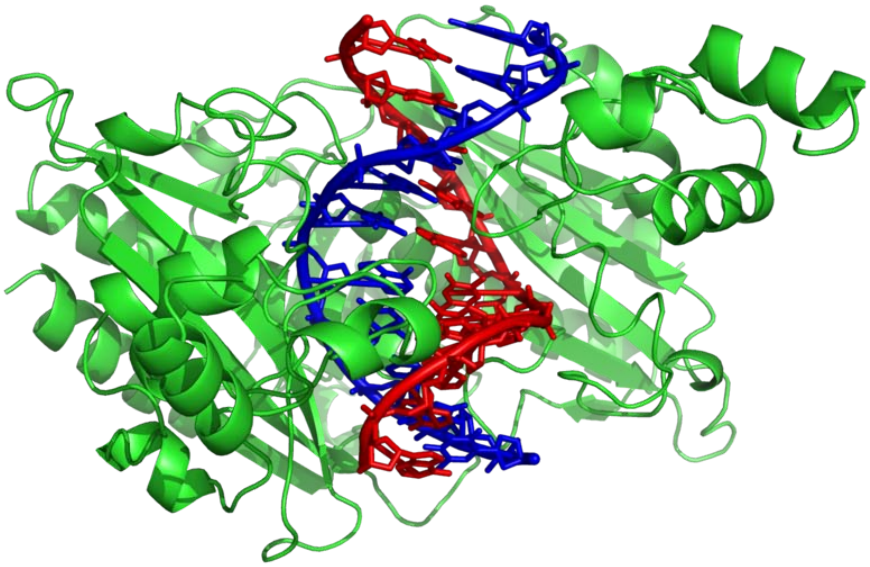
A distinct group of DNA-binding proteins are the DNA-binding proteins that specifically bind single-stranded DNA. In humans, replication protein A is the best-understood member of this family and is used in processes where the double helix is separated, including DNA replication, recombination and DNA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



The lambda repressor helix-turn-helix transcription factor bound to its DNA target

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 will change 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.



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

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. 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

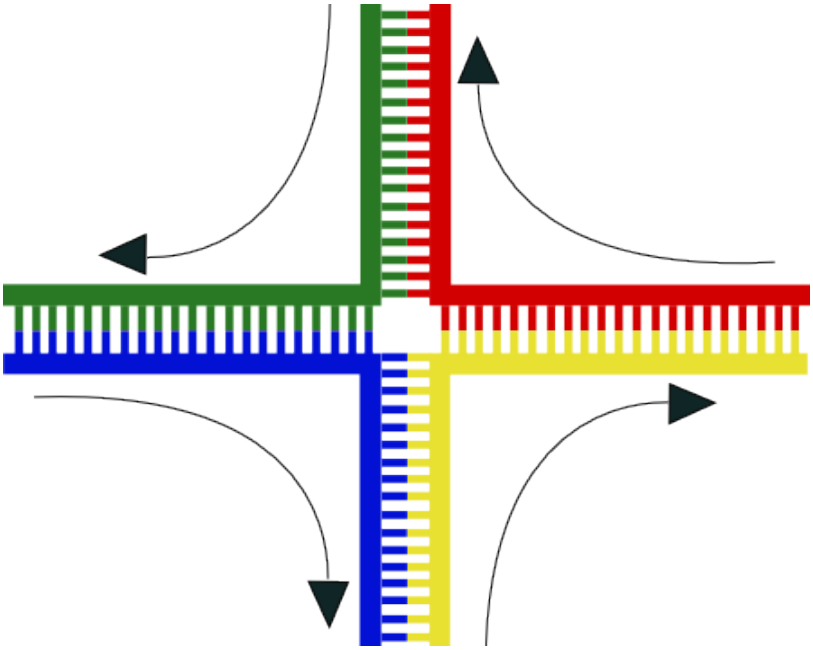
Polymerases are enzymes that synthesize polynucleotide chains from nucleoside triphosphates. The sequence of their products are copies of existing polynucleotide chains - which are called *templates*. These enzymes function by adding nucleotides onto the 3′ hydroxyl group of the previous nucleotide in a DNA strand. 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.

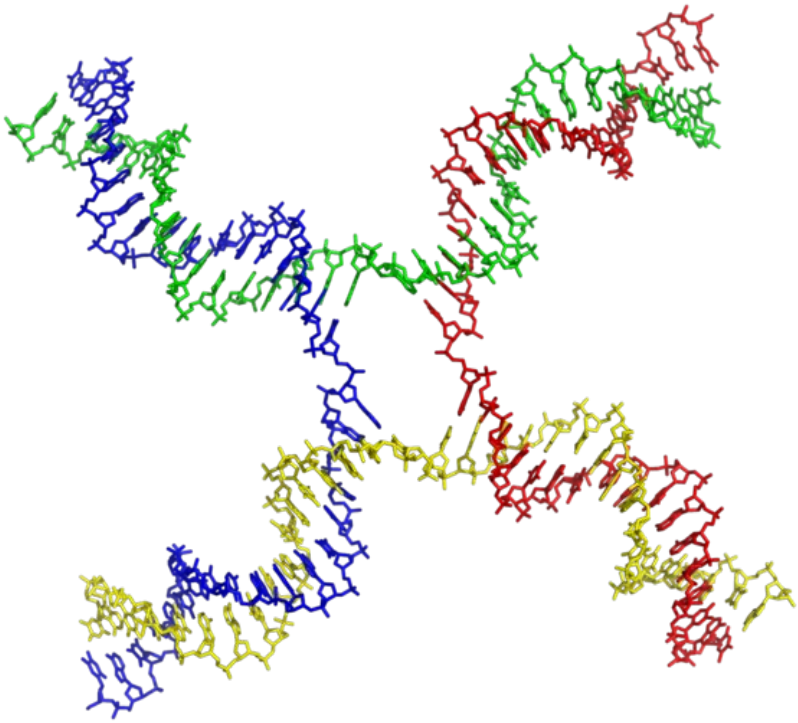
In DNA replication, a DNA-dependent DNA polymerase makes a copy of a DNA sequence. Accuracy is vital in this process, so many of these polymerases have a proofreading activity. Here, the polymerase recognizes the occasional mistakes in the synthesis reaction by the lack of base pairing between the mismatched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

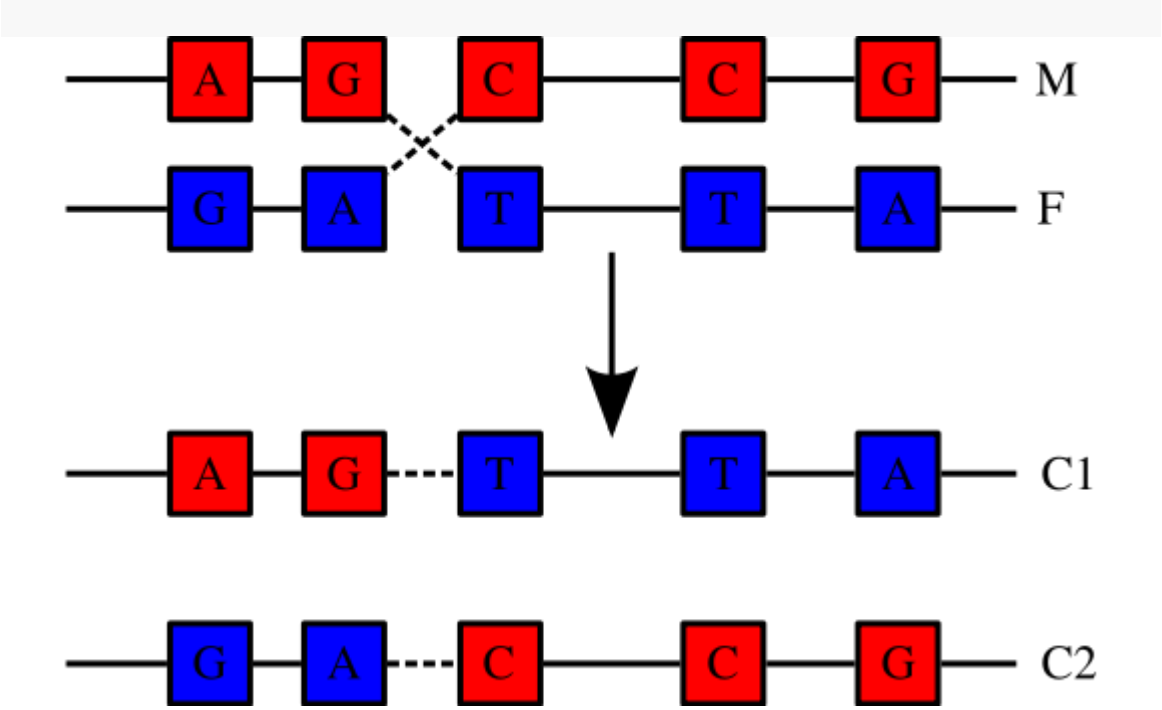
Transcription is carried out by a DNA-dependent RNA polymerase that copies the sequence of a DNA strand into RNA. To begin transcribing a gene, the RNA polymerase binds to a sequence of DNA called a promoter and separates the DNA strands. It then copies the gene sequence into a messenger RNA transcript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





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



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". 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. 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 either caused by 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. This is because DNA will survive 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.

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.

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 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

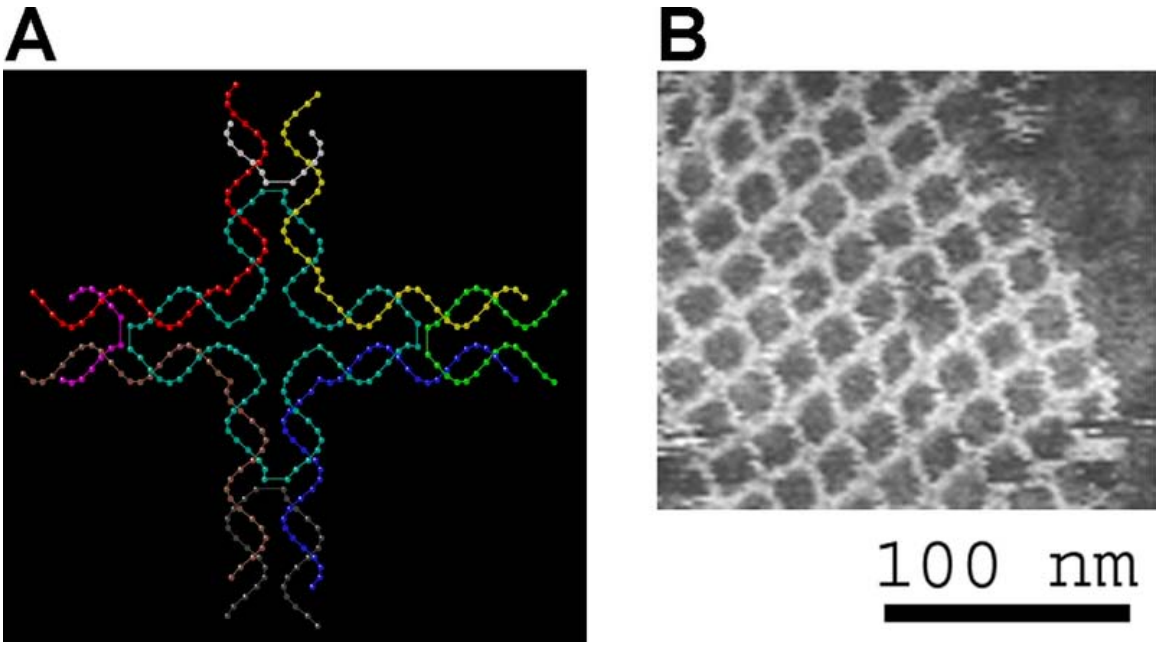
was developed in 1984 by British geneticist Sir Alec Jeffreys, and first used in forensic science to convict Colin Pitchfork in the 1988 Enderby murders case.

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. 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 biological data, and this includes 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. 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 sequenced 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



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 that seeks to design nanoscale structures using the molecular recognition properties of DNA molecules.

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 as well as 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.

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.

History of DNA research

James D. Watson and Francis Crick (right), co-originators of the double-helix model, with Maclyn McCarty (left)

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 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. 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.

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

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. 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 D. 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 — also obtained through private communications from Erwin Chargaff in the previous years. Chargaff's rules played a

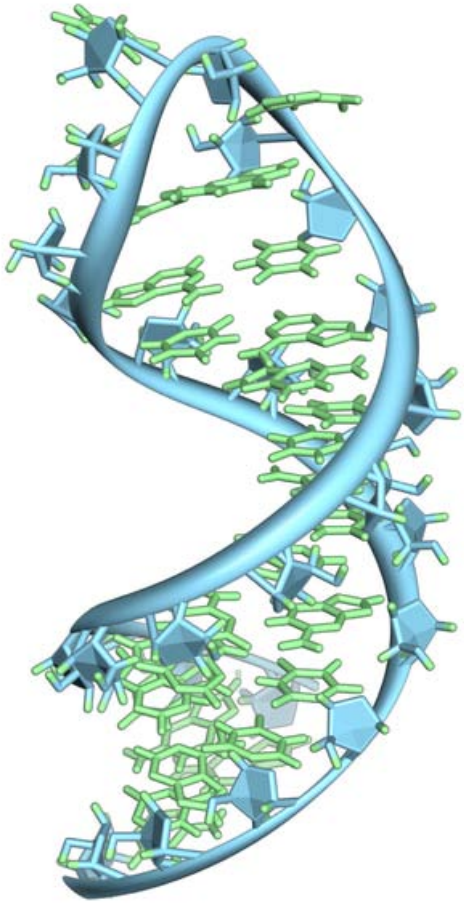
very important role in establishing double-helix configurations for B-DNA as well as A-DNA.

Experimental evidence supporting the Watson and Crick model were published in a series of five articles in the same issue of *Nature*. 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. However, Nobel rules of the time allowed only living recipients, but a vigorous debate continues on 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.

Chapter- 3

RNA



A hairpin loop from a pre-mRNA. Highlighted are the nucleobases (green) and the ribose-phosphate backbone (blue).

Ribonucleic acid (**RNA**) is one of the three major macromolecules (along with DNA and proteins) that are essential for all known forms of life.

Like DNA, RNA is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase (sometimes called a nitrogenous base), a ribose sugar,

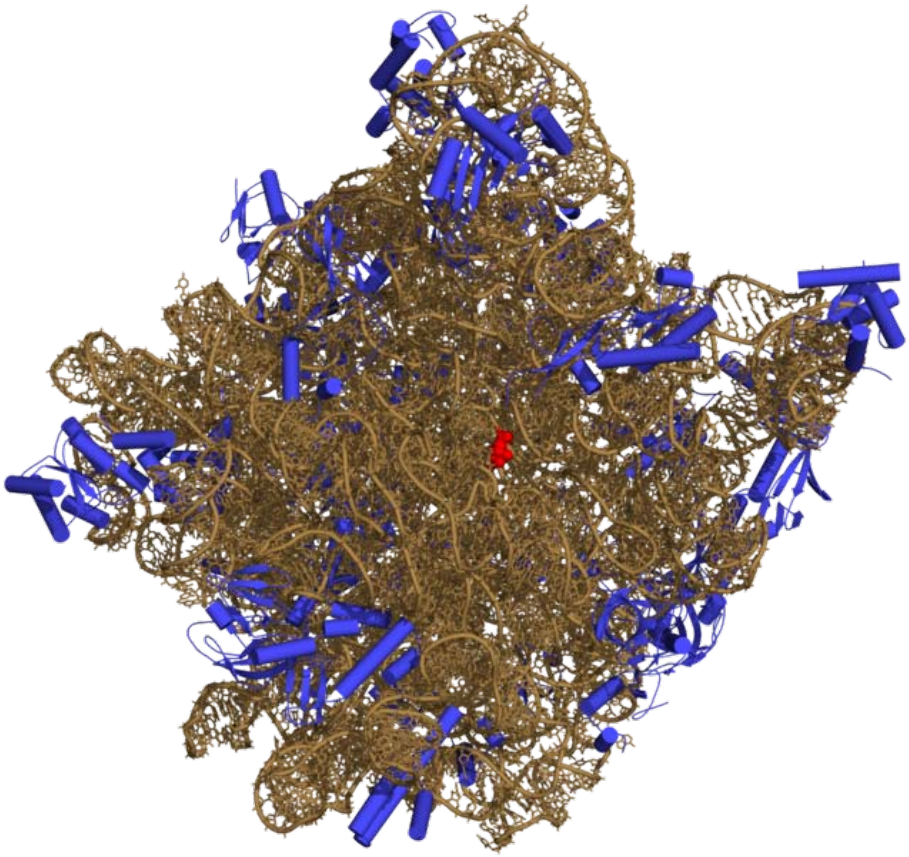
and a phosphate group. The sequence of nucleotides allows RNA to encode genetic information. For example, some viruses use RNA instead of DNA as their genetic material, and all organisms use messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins.

Like proteins, some RNA molecules play an active role in cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. One of these active processes is protein synthesis, a universal function whereby mRNA molecules direct the assembly of proteins on ribosomes. This process uses transfer RNA (tRNA) molecules to deliver amino acids to the ribosome, where ribosomal RNA (rRNA) links amino acids together to form proteins.

The chemical structure of RNA is very similar to that of DNA, with two differences--(a) RNA contains the sugar ribose while DNA contains the slightly different sugar deoxyribose (a type of ribose that lacks one oxygen atom), and (b) RNA has the nucleobase uracil while DNA contains thymine (uracil and thymine have similar base-pairing properties).

Unlike DNA, most RNA molecules are single-stranded. Single-stranded RNA molecules adopt very complex three-dimensional structures, since they are not restricted to the repetitive double-helical form of double-stranded DNA. RNA is made within living cells by RNA polymerases, enzymes that act to copy a DNA or RNA template into a new RNA strand through processes known as transcription or RNA replication, respectively.

Comparison with DNA

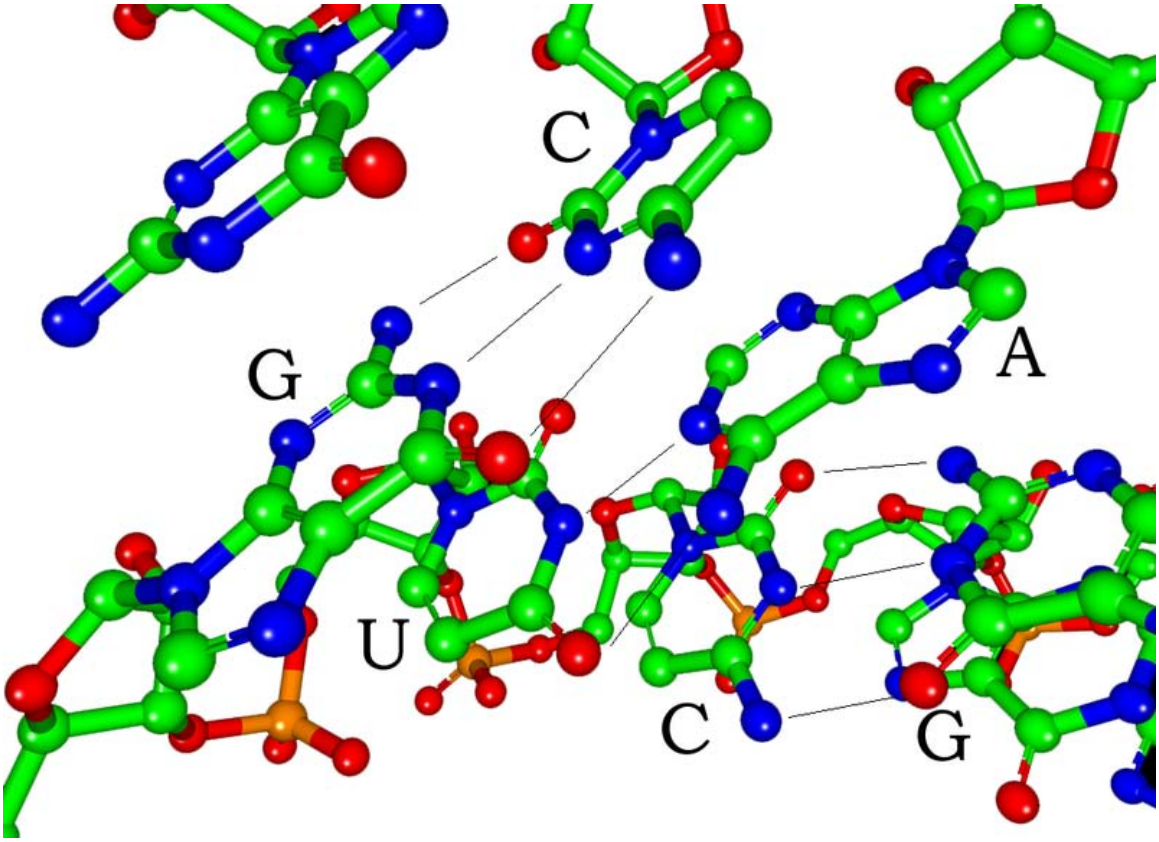


Three-dimensional representation of the 50S ribosomal subunit. RNA is in ochre, protein in blue. The active site is in the middle (red).

RNA and DNA are both nucleic acids, but differ in three main ways. First, unlike DNA, which is, in general, double-stranded, RNA is a single-stranded molecule in many of its biological roles and has a much shorter chain of nucleotides. Second, while DNA contains *deoxyribose*, RNA contains *ribose* (in deoxyribose there is no hydroxyl group attached to the pentose ring in the 2' position). These hydroxyl groups make RNA less stable than DNA because it is more prone to hydrolysis. Third, the complementary base to adenine is not thymine, as it is in DNA, but rather uracil, which is an unmethylated form of thymine.

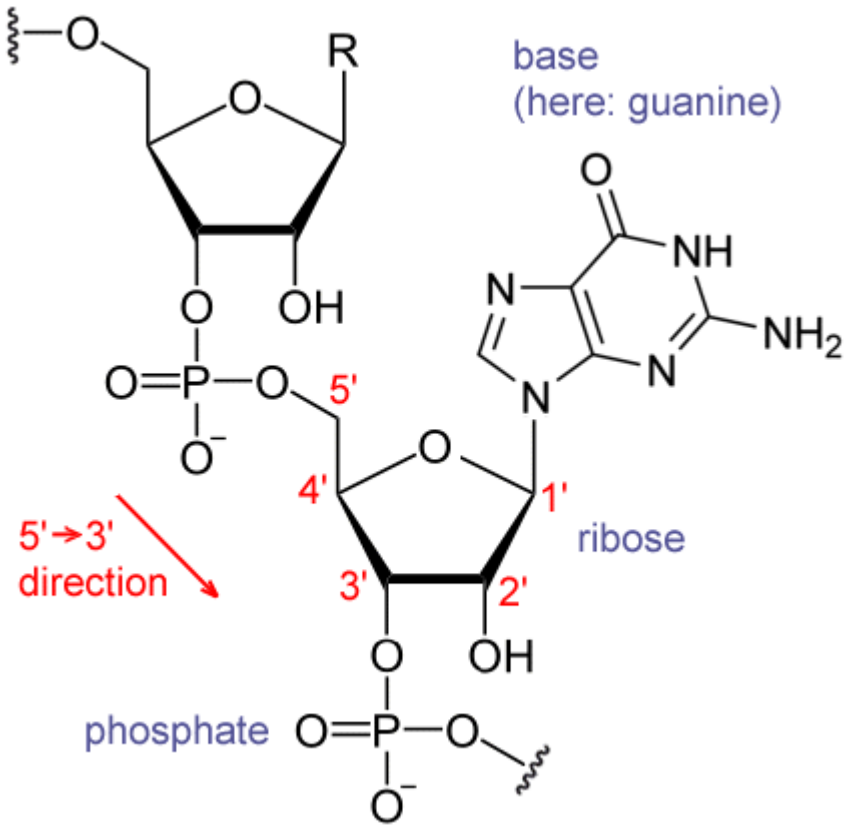
Like DNA, most biologically active RNAs, including mRNA, tRNA, rRNA, snRNAs, and other non-coding RNAs, contain self-complementary sequences that allow parts of the RNA to fold and pair with itself to form double helices. Structural analysis of these RNAs has revealed that they are highly structured. Unlike DNA, their structures do not consist of long double helices but rather collections of short helices packed together into structures akin to proteins. In this fashion, RNAs can achieve chemical catalysis, like enzymes. For instance, determination of the structure of the ribosome—an enzyme that catalyzes peptide bond formation—revealed that its active site is composed entirely of RNA.

Structure



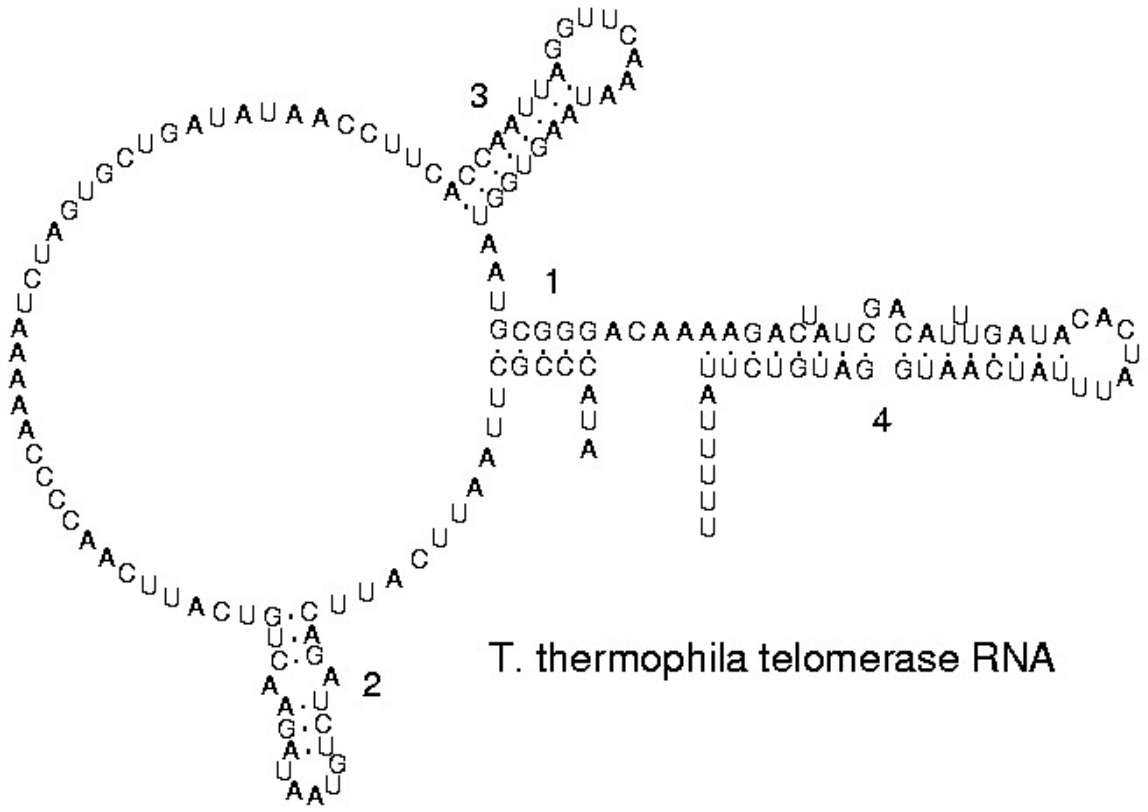
Watson-Crick base pairs in a siRNA (hydrogen atoms are not shown)

Each nucleotide in RNA contains a ribose sugar, with carbons numbered 1' through 5'. A base is attached to the 1' position, in general, adenine (A), cytosine (C), guanine (G), or uracil (U). Adenine and guanine are purines, cytosine, and uracil are pyrimidines. A phosphate group is attached to the 3' position of one ribose and the 5' position of the next. The phosphate groups have a negative charge each at physiological pH, making RNA a charged molecule (polyanion). The bases may form hydrogen bonds between cytosine and guanine, between adenine and uracil and between guanine and uracil. However, other interactions are possible, such as a group of adenine bases binding to each other in a bulge, or the GNRA tetraloop that has a guanine–adenine base-pair.



Chemical structure of RNA

An important structural feature of RNA that distinguishes it from DNA is the presence of a hydroxyl group at the 2' position of the ribose sugar. The presence of this functional group causes the helix to adopt the A-form geometry rather than the B-form most commonly observed in DNA. This results in a very deep and narrow major groove and a shallow and wide minor groove. A second consequence of the presence of the 2'-hydroxyl group is that in conformationally flexible regions of an RNA molecule (that is, not involved in formation of a double helix), it can chemically attack the adjacent phosphodiester bond to cleave the backbone.



Secondary structure of a telomerase RNA

RNA is transcribed with only four bases (adenine, cytosine, guanine and uracil), but these bases and attached sugars can be modified in numerous ways as the RNAs mature. Pseudouridine (Ψ), in which the linkage between uracil and ribose is changed from a C–N bond to a C–C bond, and ribothymidine (T) are found in various places (the most notable ones being in the TΨC loop of tRNA). Another notable modified base is hypoxanthine, a deaminated adenine base whose nucleoside is called inosine (I). Inosine plays a key role in the wobble hypothesis of the genetic code.

There are nearly 100 other naturally occurring modified nucleosides, of which pseudouridine and nucleosides with 2'-O-methylribose are the most common. The specific roles of many of these modifications in RNA are not fully understood. However, it is notable that, in ribosomal RNA, many of the post-transcriptional modifications occur in highly functional regions, such as the peptidyl transferase center and the subunit interface, implying that they are important for normal function.

The functional form of single stranded RNA molecules, just like proteins, frequently requires a specific tertiary structure. The scaffold for this structure is provided by secondary structural elements that are hydrogen bonds within the molecule. This leads to several recognizable "domains" of secondary structure like hairpin loops, bulges, and

internal loops. Since RNA is charged, metal ions such as Mg^{2+} are needed to stabilise many secondary and tertiary structures.

Synthesis

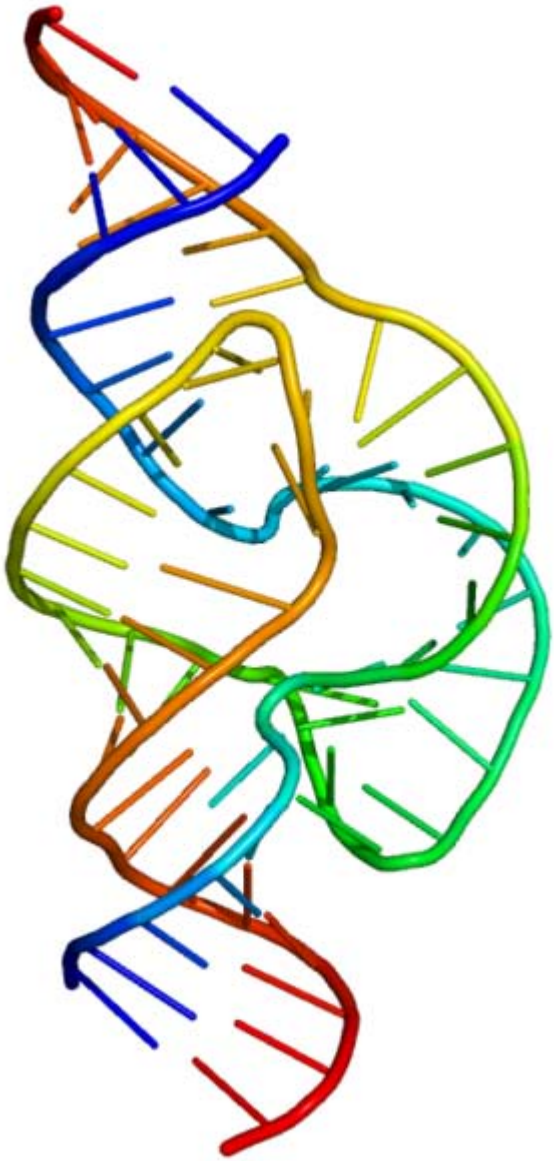
Synthesis of RNA is usually catalyzed by an enzyme—RNA polymerase—using DNA as a template, a process known as transcription. Initiation of transcription begins with the binding of the enzyme to a promoter sequence in the DNA (usually found "upstream" of a gene). The DNA double helix is unwound by the helicase activity of the enzyme. The enzyme then progresses along the template strand in the 3’ to 5’ direction, synthesizing a complementary RNA molecule with elongation occurring in the 5’ to 3’ direction. The DNA sequence also dictates where termination of RNA synthesis will occur.

RNAs are often modified by enzymes after transcription. For example, a poly(A) tail and a 5' cap are added to eukaryotic pre-mRNA and introns are removed by the spliceosome.

There are also a number of RNA-dependent RNA polymerases that use RNA as their template for synthesis of a new strand of RNA. For instance, a number of RNA viruses (such as poliovirus) use this type of enzyme to replicate their genetic material. Also, RNA-dependent RNA polymerase is part of the RNA interference pathway in many organisms.

Types of RNA

Overview



Structure of a hammerhead ribozyme, a ribozyme that cuts RNA

Messenger RNA (mRNA) is the RNA that carries information from DNA to the ribosome, the sites of protein synthesis (translation) in the cell. The coding sequence of the mRNA determines the amino acid sequence in the protein that is produced. Many RNAs do not code for protein however (about 97% of the transcriptional output is non-protein-coding in eukaryotes).

These so-called non-coding RNAs ("ncRNA") can be encoded by their own genes (RNA genes), but can also derive from mRNA introns. The most prominent examples of non-coding RNAs are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation. There are also non-coding RNAs involved in gene regulation, RNA processing and other roles. Certain RNAs are able to catalyse chemical reactions such as cutting and ligating other RNA molecules, and the catalysis of peptide bond formation in the ribosome; these are known as ribozymes.

In translation

Messenger RNA (mRNA) carries information about a protein sequence to the ribosomes, the protein synthesis factories in the cell. It is coded so that every three nucleotides (a codon) correspond to one amino acid. In eukaryotic cells, once precursor mRNA (pre-mRNA) has been transcribed from DNA, it is processed to mature mRNA. This removes its introns—non-coding sections of the pre-mRNA. The mRNA is then exported from the nucleus to the cytoplasm, where it is bound to ribosomes and translated into its corresponding protein form with the help of tRNA. In prokaryotic cells, which do not have nucleus and cytoplasm compartments, mRNA can bind to ribosomes while it is being transcribed from DNA. After a certain amount of time the message degrades into its component nucleotides with the assistance of ribonucleases.

Transfer RNA (tRNA) is a small RNA chain of about 80 nucleotides that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. It has sites for amino acid attachment and an anticodon region for codon recognition that binds to a specific sequence on the messenger RNA chain through hydrogen bonding.

Ribosomal RNA (rRNA) is the catalytic component of the ribosomes. Eukaryotic ribosomes contain four different rRNA molecules: 18S, 5.8S, 28S and 5S rRNA. Three of the rRNA molecules are synthesized in the nucleolus, and one is synthesized elsewhere. In the cytoplasm, ribosomal RNA and protein combine to form a nucleoprotein called a ribosome. The ribosome binds mRNA and carries out protein synthesis. Several ribosomes may be attached to a single mRNA at any time. rRNA is extremely abundant and makes up 80% of the 10 mg/ml RNA found in a typical eukaryotic cytoplasm.

Transfer-messenger RNA (tmRNA) is found in many bacteria and plastids. It tags proteins encoded by mRNAs that lack stop codons for degradation and prevents the ribosome from stalling.

Regulatory RNAs

Several types of RNA can downregulate gene expression by being complementary to a part of an mRNA or a gene's DNA. MicroRNAs (miRNA; 21-22 nt) are found in eukaryotes and act through RNA interference (RNAi), where an effector complex of miRNA and enzymes can break down mRNA to which the miRNA is complementary, block the mRNA from being translated, or accelerate its degradation. While small

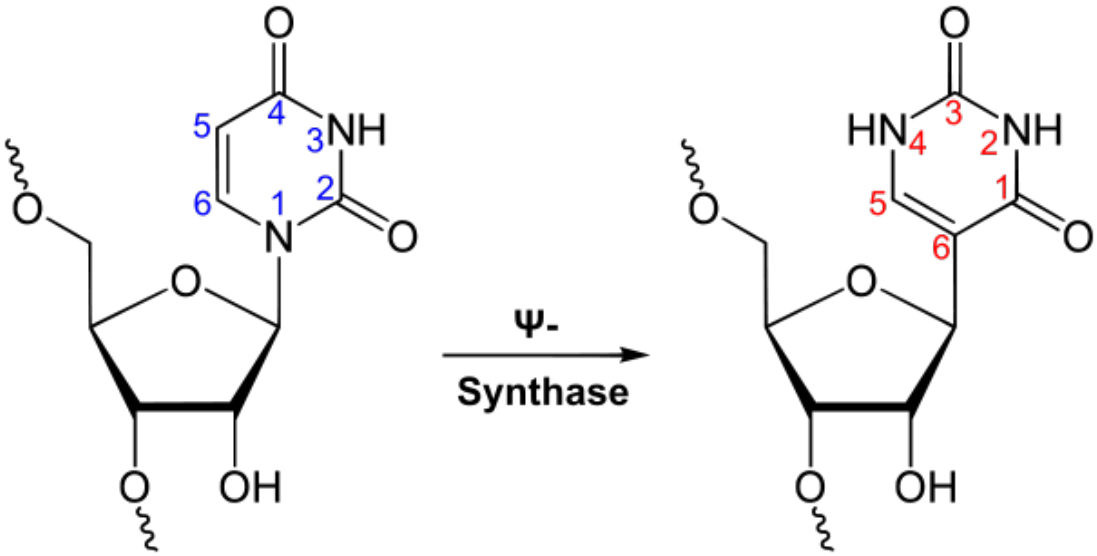
interfering RNAs (siRNA; 20-25 nt) are often produced by breakdown of viral RNA, there are also endogenous sources of siRNAs.

siRNAs act through RNA interference in a fashion similar to miRNAs. Some miRNAs and siRNAs can cause genes they target to be methylated, thereby decreasing or increasing transcription of those genes. Animals have Piwi-interacting RNAs (piRNA; 29-30 nt) which are active in germline cells and are thought to be a defense against transposons and play a role in gametogenesis.

Many prokaryotes have CRISPR RNAs, a regulatory system similar to RNA interference. Antisense RNAs are widespread; most downregulate a gene, but a few are activators of transcription. One way antisense RNA can act is by binding to an mRNA, forming double-stranded RNA that is enzymatically degraded. There are many long noncoding RNAs that regulate genes in eukaryotes, one such RNA is Xist, which coats one X chromosome in female mammals and inactivates it.

An mRNA may contain regulatory elements itself, such as riboswitches, in the 5' untranslated region or 3' untranslated region; these cis-regulatory elements regulate the activity of that mRNA. The untranslated regions can also contain elements that regulate other genes.

In RNA processing



Uridine to pseudouridine is a common RNA modification.

Many RNAs are involved in modifying other RNAs. Introns are spliced out of pre-mRNA by spliceosomes, which contain several small nuclear RNAs (snRNA), or the introns can be ribozymes that are spliced by themselves. RNA can also be altered by having its nucleotides modified to other nucleotides than A, C, G and U. In eukaryotes, modifications of RNA nucleotides are generally directed by small nucleolar RNAs

(snoRNA; 60-300 nt), found in the nucleolus and cajal bodies. snoRNAs associate with enzymes and guide them to a spot on an RNA by basepairing to that RNA. These enzymes then perform the nucleotide modification. rRNAs and tRNAs are extensively modified, but snRNAs and mRNAs can also be the target of base modification.

RNA genomes

Like DNA, RNA can carry genetic information. RNA viruses have genomes composed of RNA, and a variety of proteins encoded by that genome. The viral genome is replicated by some of those proteins, while other proteins protect the genome as the virus particle moves to a new host cell. Viroids are another group of pathogens, but they consist only of RNA, do not encode any protein and are replicated by a host plant cell's polymerase.

In reverse transcription

Reverse transcribing viruses replicate their genomes by reverse transcribing DNA copies from their RNA; these DNA copies are then transcribed to new RNA. Retrotransposons also spread by copying DNA and RNA from one another, and telomerase contains an RNA that is used as template for building the ends of eukaryotic chromosomes.

Double-stranded RNA

Double-stranded RNA (dsRNA) is RNA with two complementary strands, similar to the DNA found in all cells. dsRNA forms the genetic material of some viruses (double-stranded RNA viruses). Double-stranded RNA such as viral RNA or siRNA can trigger RNA interference in eukaryotes, as well as interferon response in vertebrates.

Key discoveries in RNA biology

Research on RNA has led to many important biological discoveries and numerous Nobel Prizes. Nucleic acids were discovered in 1868 by Friedrich Miescher, who called the material 'nuclein' since it was found in the nucleus. It was later discovered that prokaryotic cells, which do not have a nucleus, also contain nucleic acids. The role of RNA in protein synthesis was suspected already in 1939. Severo Ochoa won the 1959 Nobel Prize in Medicine (shared with Arthur Kornberg) after he discovered an enzyme that can synthesize RNA in the laboratory. Ironically, the enzyme discovered by Ochoa (polynucleotide phosphorylase) was later shown to be responsible for RNA degradation, not RNA synthesis.

The sequence of the 77 nucleotides of a yeast tRNA was found by Robert W. Holley in 1965, winning Holley the 1968 Nobel Prize in Medicine (shared with Har Gobind Khorana and Marshall Nirenberg). In 1967, Carl Woese hypothesized that RNA might be catalytic and suggested that the earliest forms of life (self-replicating molecules) could have relied on RNA both to carry genetic information and to catalyze biochemical reactions—an RNA world.

During the early 1970s, retroviruses and reverse transcriptase were discovered, showing for the first time that enzymes could copy RNA into DNA (the opposite of the usual route for transmission of genetic information). For this work, David Baltimore, Renato Dulbecco and Howard Temin were awarded a Nobel Prize in 1975. In 1976, Walter Fiers and his team determined the first complete nucleotide sequence of an RNA virus genome, that of bacteriophage MS2.

In 1977, introns and RNA splicing were discovered in both mammalian viruses and in cellular genes, resulting in a 1993 Nobel to Philip Sharp and Richard Roberts. Catalytic RNA molecules (ribozymes) were discovered in the early 1980s, leading to a 1989 Nobel award to Thomas Cech and Sidney Altman. In 1990 it was found in petunia that introduced genes can silence similar genes of the plant's own, now known to be a result of RNA interference.

At about the same time, 22 nt long RNAs, now called microRNAs, were found to have a role in the development of *C. elegans*. Studies on RNA interference gleaned a Nobel Prize for Andrew Fire and Craig Mello in 2006, and another Nobel was awarded for studies on transcription of RNA to Roger Kornberg in the same year. The discovery of gene regulatory RNAs has led to attempts to develop drugs made of RNA, such as siRNA, to silence genes.

Chapter- 4

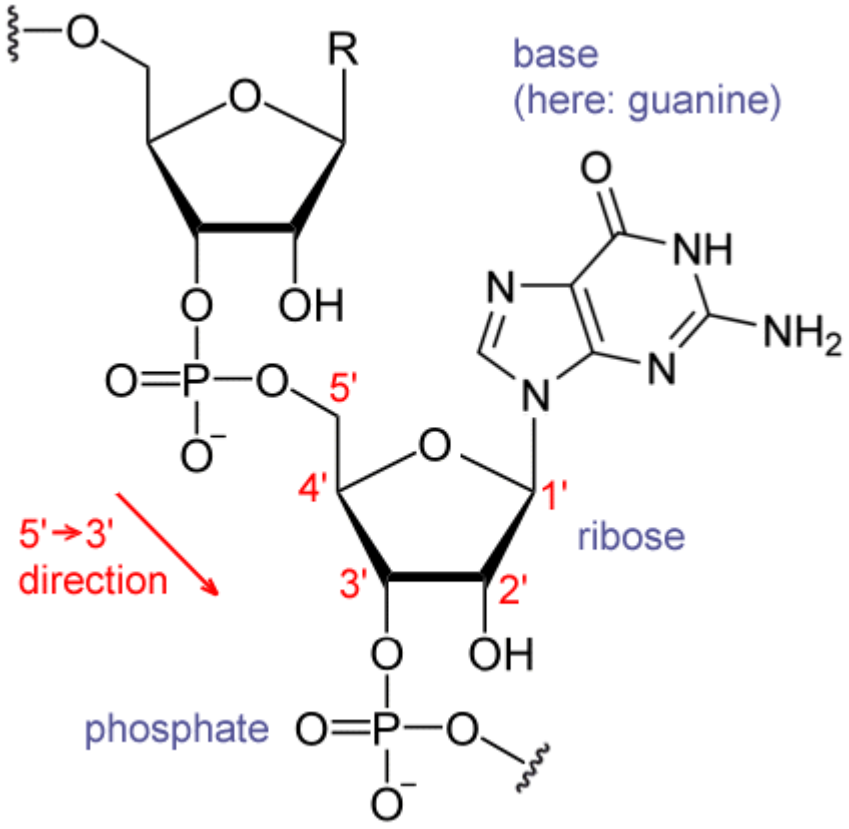
Nucleic Acid Sequence

The **sequence** or **primary structure of a nucleic acid** is the exact specification of its atomic composition and the chemical bonds connecting those atoms. As nucleic acids, e.g. DNA and RNA, are unbranched polymers, this is equivalent to specifying exact sequence of nucleotides that comprise the whole molecule. This sequence is written as a succession of letters representing a real or hypothetical DNA molecule or strand. By convention, the primary structure of a DNA or RNA molecule is reported from the 5' end to the 3' end.

The sequence has capacity to carry information. When used in reference to biological DNA, which carries the information which directs the functions of living beings, the term **genetic sequence** is often used. Sequences can be read from the biological raw material through DNA sequencing methods.

Primary structure is sometimes mistakenly termed *primary sequence*, but there is no such term, as well as no parallel concept of secondary or tertiary sequence.

Nucleotides



Chemical structure of RNA

Nucleic acids consist of a chain of linked units called nucleotides. Each nucleotide consists of three subunits: a phosphate group and a sugar (ribose in the case of RNA, deoxyribose in DNA) make up the backbone of the nucleic acid strand, and attached to the sugar is one of a set of nucleobases. The nucleobases are important in base pairing of strands to form higher-level secondary and tertiary structure such as the famed double helix.

The possible letters are *A*, *C*, *G*, and *T*, representing the four nucleotide bases of a DNA strand — adenine, cytosine, guanine, thymine — covalently linked to a phosphodiester backbone. In the typical case, the sequences are printed abutting one another without gaps, as in the sequence AAAGTCTGAC, read left to right in the 5' to 3' direction. With regards to transcription, a sequence is on the coding strand if it has the same order as the transcribed RNA.

One sequence can be complementary to another sequence, meaning that they have the base on each position is the complementary (i.e. A to T, C to G) and in the reverse order. For example, the complementary sequence to TTAC is GTAA. If one strand of the double-stranded DNA is considered the sense strand, then the other strand, considered the antisense strand, will have the complementary sequence to the sense strand.

Notation

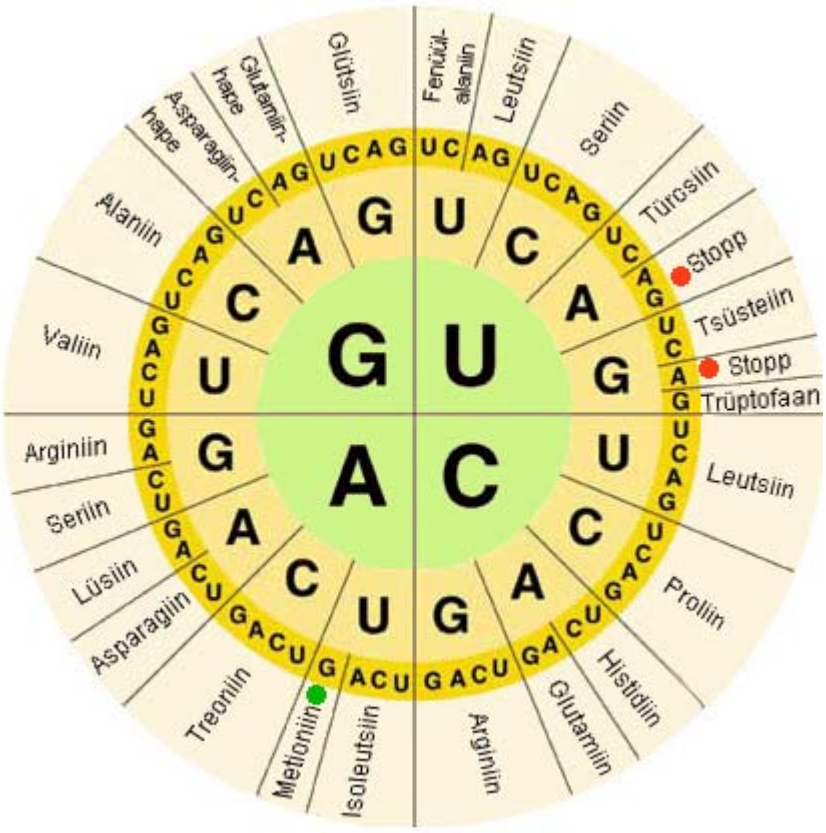
While A, T, C, and G represent a particular nucleotide at a position, there are also letters that represent ambiguity. Of all the molecules sampled, there is more than one kind of nucleotide at that position. The rules of the International Union of Pure and Applied Chemistry (IUPAC) are as follows:

- **A** = adenine
- **C** = cytosine
- **G** = guanine
- **T** = thymine
- **R** = G A (purine)
- **Y** = T C (pyrimidine)
- **K** = G T (keto)
- **M** = A C (amino)
- **S** = G C (strong bonds)
- **W** = A T (weak bonds)
- **B** = G T C (all but A)
- **D** = G A T (all but C)
- **H** = A C T (all but G)
- **V** = G C A (all but T)
- **N** = A G C T (any)

These symbols are also valid for RNA, except with U (uracil) replacing T (thymine).

Apart from adenine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), DNA and RNA also contain bases that have been modified after the nucleic acid chain has been formed. In DNA, the most common modified base is 5-methylcytidine (m5C). In RNA, there are many modified bases, including pseudouridine (Ψ), dihydrouridine (D), inosine (I), ribothymidine (rT) and 7-methylguanosine (m7G). Hypoxanthine and xanthine are two of the many bases created through mutagen presence, both of them through deamination (replacement of the amine-group with a carbonyl-group). Hypoxanthine is produced from adenine, xanthine from guanine. Similarly, deamination of cytosine results in uracil.

Biological significance

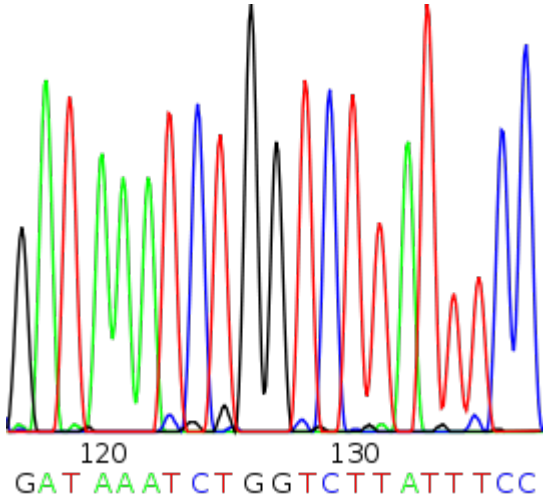


A depiction of the genetic code, by which the information contained in nucleic acids are translated into amino acid sequences in proteins.

In biological systems, nucleic acids contain information which is used by a living cell to construct specific proteins. The sequence of nucleobases on a nucleic acid strand is translated by cell machinery into a sequence of amino acids making up a protein strand. Each group of three bases, called a codon, corresponds to a single amino acid, and there is a specific genetic code by which each possible combination of three bases corresponds to a specific amino acid.

The central dogma of molecular biology outlines the mechanism by which proteins are constructed using information contained in nucleic acids. DNA is transcribed into mRNA molecules, which travels to the ribosome where the mRNA is used as a template for the construction of the protein strand. Since nucleic acids can bind to molecules with complementary sequences, there is a distinction between "sense" sequences which code for proteins, and the complementary "antisense" sequence which is by itself nonfunctional, but can bind to the sense strand.

Sequence determination



Electropherogram printout from automated sequencer for determining part of a DNA sequence

DNA sequencing is the process of determining the nucleotide sequence of a given DNA fragment. The sequence of DNA encodes the necessary information for living things to survive and reproduce. Determining the sequence is therefore useful in fundamental research into why and how organisms live, as well as in applied subjects. Because of the key nature of DNA to living things, knowledge of DNA sequence may come in useful in practically any biological research. For example, in medicine it can be used to identify, diagnose and potentially develop treatments for genetic diseases. Similarly, research into pathogens may lead to treatments for contagious diseases. Biotechnology is a burgeoning discipline, with the potential for many useful products and services.

RNA is not sequenced directly. Instead, it is copied to a DNA by reverse transcriptase, and this DNA is then sequenced.

Current sequencing methods rely on the discriminatory ability of DNA polymerases, and can therefore only distinguish four bases. An inosine (created from adenosine during RNA editing) will be read as a G, and 5-methyl-cytosine (created from cytosine by DNA methylation) will be read as a C. It is also currently difficult to sequence small amounts of DNA, as the signal will be too weak to measure. This is overcome by PCR amplification.

Digital format

Genetic sequence in digital format

Once a nucleic acid sequence has been obtained from an organism, it is stored *in silico* in digital format. Digital genetic sequences may be stored in sequence databases, be analyzed, be digitally altered and/or be used as templates for creating new actual DNA using artificial gene synthesis.

Sequence analysis

Digital genetic sequences may be analyzed using the tools of bioinformatics to attempt to determine its function.

Genetic testing

The DNA in an organism's genome can be analyzed to diagnose vulnerabilities to inherited diseases, and can also be used to determine a child's paternity (genetic father) or a person's ancestry. Normally, every person carries two copies of every gene, one inherited from their mother, one inherited from their father. The human genome is believed to contain around 20,000 - 25,000 genes. In addition to studying chromosomes to the level of individual genes, genetic testing in a broader sense includes biochemical tests for the possible presence of genetic diseases, or mutant forms of genes associated with increased risk of developing genetic disorders.

Genetic testing identifies changes in chromosomes, genes, or proteins. Most of the time, testing is used to find changes that are associated with inherited disorders. The results of a genetic test can confirm or rule out a suspected genetic condition or help determine a person's chance of developing or passing on a genetic disorder. Several hundred genetic tests are currently in use, and more are being developed.

Sequence alignment

In bioinformatics, a **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In sequence alignments of proteins, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages. The absence of substitutions, or the presence of only very conservative substitutions (that is, the substitution of amino acids whose side chains have similar biochemical properties) in a particular region of the sequence, suggest that this region has structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than are amino acids, the conservation of base pairs can indicate a similar functional or structural role.

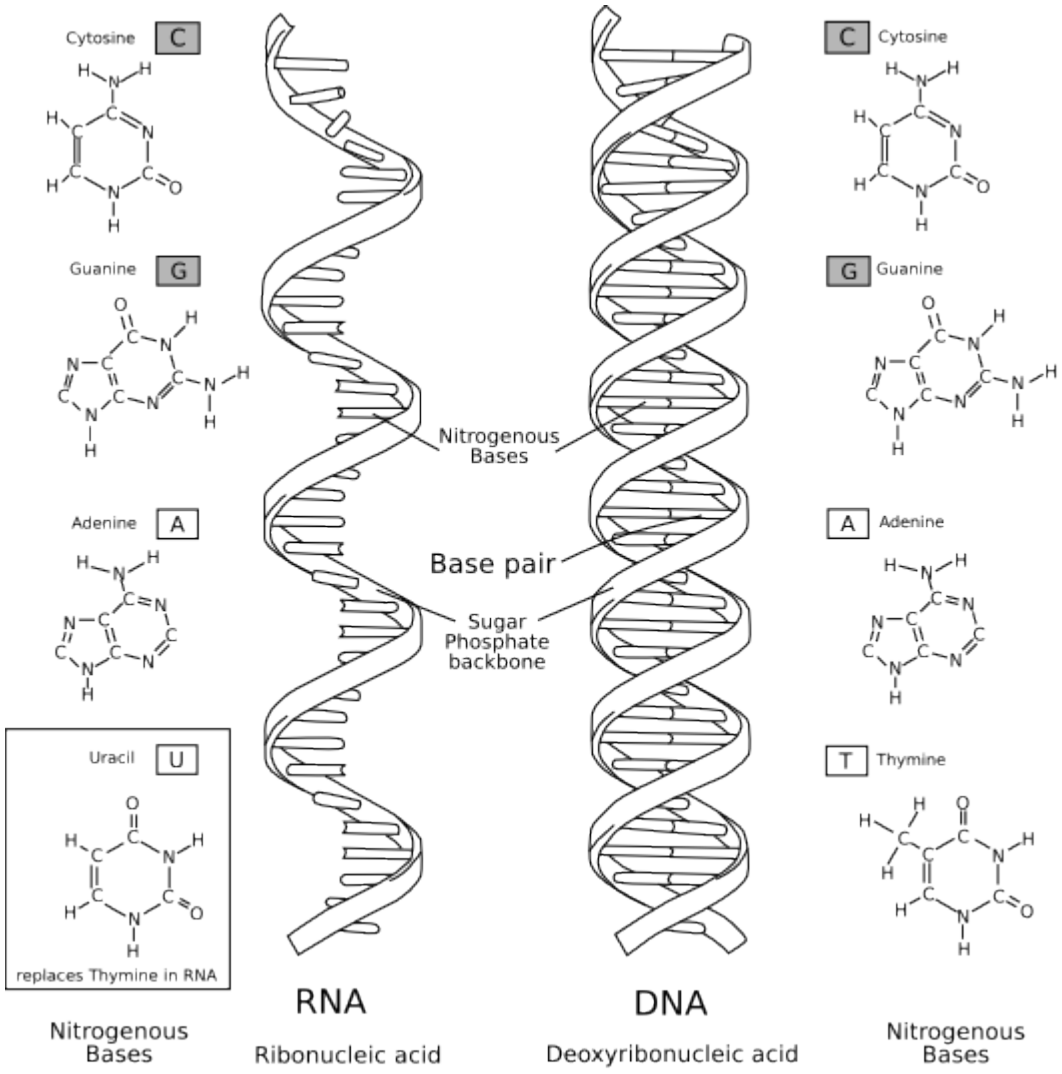
Computational phylogenetics makes extensive use of sequence alignments in the construction and interpretation of phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species. The degree to which sequences in a query set differ is qualitatively related to the sequences' evolutionary distance from one another. Roughly speaking, high sequence identity suggests that the sequences in question have a comparatively young most recent common ancestor, while low identity suggests that the divergence is more ancient. This approximation, which reflects the "molecular clock" hypothesis that a roughly constant rate of evolutionary change can be used to extrapolate the elapsed time since two genes first diverged (that is, the coalescence time), assumes that the effects of mutation and selection are constant across sequence lineages. Therefore it does not account for possible difference among organisms or species in the rates of DNA repair or the possible functional conservation of specific regions in a sequence. (In the case of nucleotide sequences, the molecular clock hypothesis in its most basic form also discounts the difference in acceptance rates between silent mutations that do not alter the meaning of a given codon and other mutations that result in a different amino acid being incorporated into the protein.) More statistically accurate methods allow the evolutionary rate on each branch of the phylogenetic tree to vary, thus producing better estimates of coalescence times for genes.

Sequence motifs

Frequently the primary structure encodes motifs that are of functional importance. Some examples of sequence motifs are: the C/D and H/ACA boxes of snoRNAs, Sm binding site found in spliceosomal RNAs such as U1, U2, U4, U5, U6, U12 and U3, the Shine-Dalgarno sequence, the Kozak consensus sequence and the RNA polymerase III terminator.

Chapter- 5

Nucleic Acid Analogues



RNA with its nucleobases to the left and DNA to the right

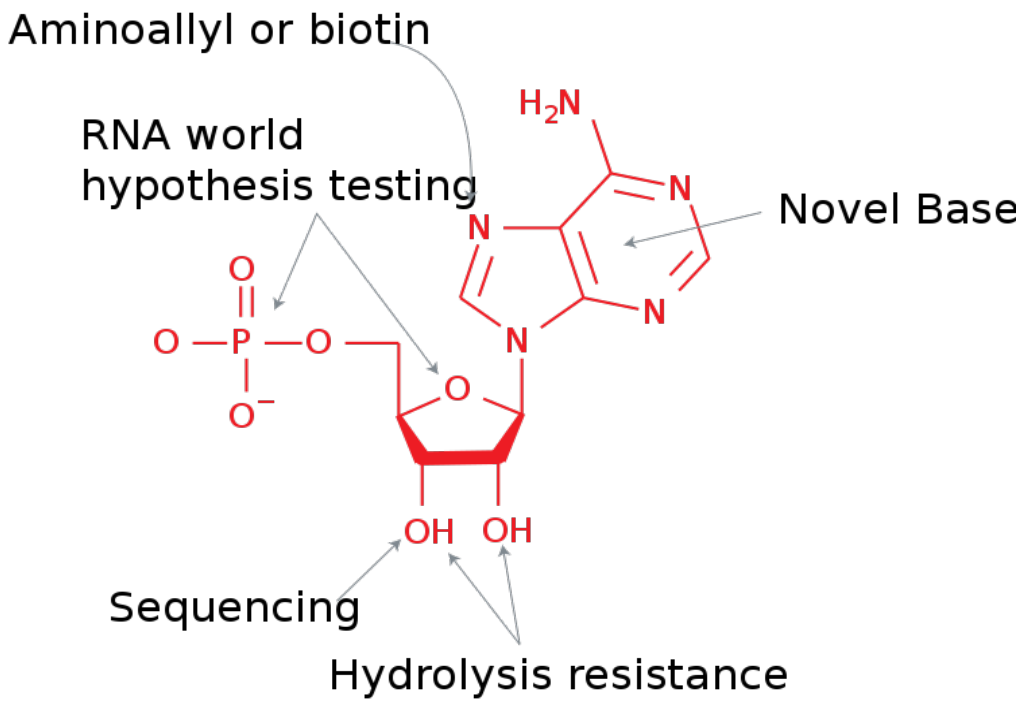
Nucleic acid analogues are compounds structurally similar (analog) to naturally occurring RNA and DNA, used in medicine and in molecular biology research. Nucleic acids are chains of nucleotides, which are composed of three parts: a phosphate backbone, a pucker-shaped pentose sugar, either ribose or deoxyribose, and one of four nucleobases. An analogue may have any of these altered. Typically the analogue nucleobases confer, among other things, different base pairing and base stacking proprieties. Examples include universal bases, which can pair with all four canon bases, and phosphate-sugar backbone analogues such as PNA, which affect the properties of the chain (PNA can even form a triple helix).

Artificial nucleic acids include peptide nucleic acid (PNA), Morpholino and locked nucleic acid (LNA), as well as glycol nucleic acid (GNA) and threose nucleic acid (TNA). Each of these is distinguished from naturally-occurring DNA or RNA by changes to the backbone of the molecule.

Medicine

Several nucleoside analogues are used as antiviral or anticancer agents. The viral polymerase incorporates these compounds with non-canon bases. These compounds are activated in the cells by being converted into nucleotides, they are administered as nucleosides since charged nucleotides cannot easily cross cell membranes.

Molecular biology



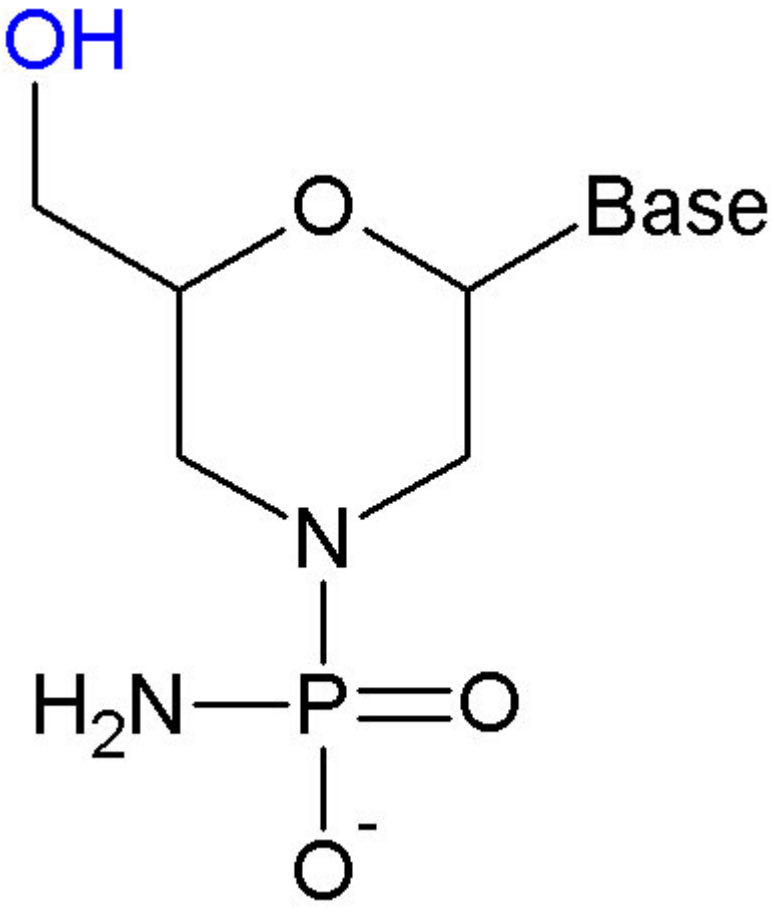
Common changes in nucleotide analogues

Nucleic acid analogues are used in molecular biology for several purposes:

- As a tool to detect particular sequences
- As a tool with resistance to RNA hydrolysis
- As a tool for another purpose, such as sequencing
- Naturally occurring, such as in tRNA
- Investigation of the mechanisms used by enzyme, such as an Enzyme inhibitor
- Investigation of possible scenarios of the origin of life
- Investigation of the structural features of nucleic acids
- Investigation of the possible alternatives to the natural system in Synthetic biology

Backbone analogues

Hydrolysis resistant RNA-analogues



Chemical structure of Morpholino

To overcome the fact that ribose's 2' hydroxy group that reacts with the phosphate linked 3' hydroxy group (RNA is too unstable to be used or synthesized reliably), a ribose analogue is used. The most common RNA analogues are locked nucleic acid (LNA),

morpholino, and peptide nucleic acid (PNA). These oligonucleotides differ as they have a different backbone sugar but still bind according to Watson and Crick pairing with RNA or DNA, but are immune to nuclease activity (They generally cannot be enzymatically synthesized and can only be produced synthetically).

Other notable analogues used as tools

Dideoxynucleotides are used in sequencing. These nucleoside triphosphates possess a non-canonical sugar, dideoxyribose, which lacks the 3' hydroxyl group normally present in DNA and therefore cannot bond with the next base. The lack of the 3' hydroxyl group terminates the chain reaction as the DNA polymerases mistake it for a regular deoxyribonucleotide. Another chain-terminating analogue that lacks a 3' hydroxyl and mimics adenosine is called cordycepin. Cordycepin is an anticancer drug that targets RNA replication. Another analogue in sequencing is a nucleobase analogue, 7-deaza-GTP and is used to sequence CG rich regions, instead 7-deaza-ATP is called tubercidin, an antibiotic.

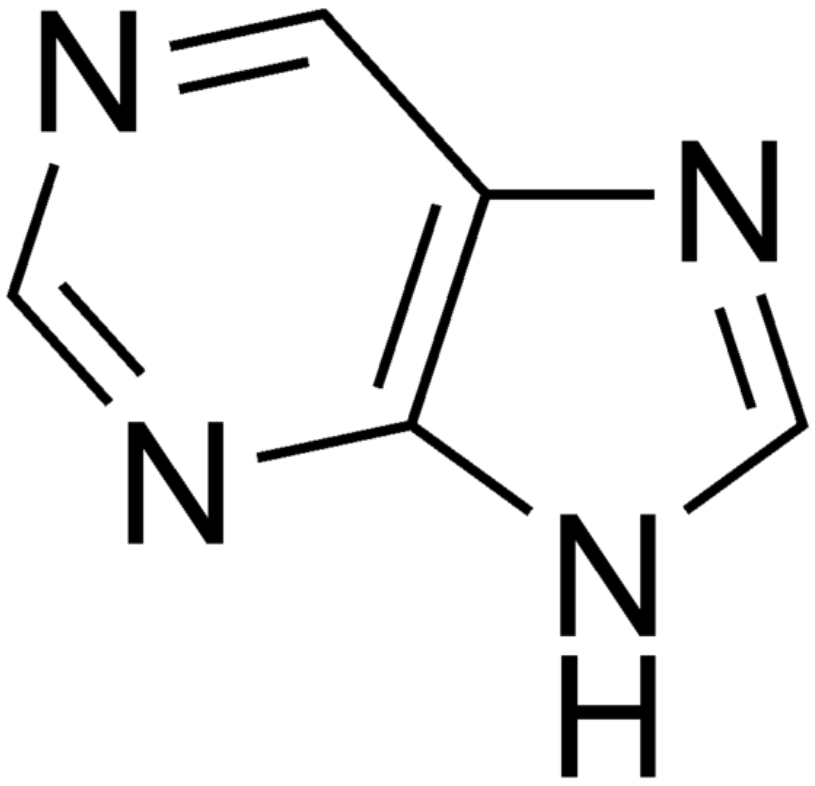
Precursors to the RNA world

RNA may be too complex to be the first nucleic acid, so before the RNA world several simpler nucleic acids that differ in the backbone, such as TNA and GNA and PNA, have been offered as candidates for the first nucleic acids.

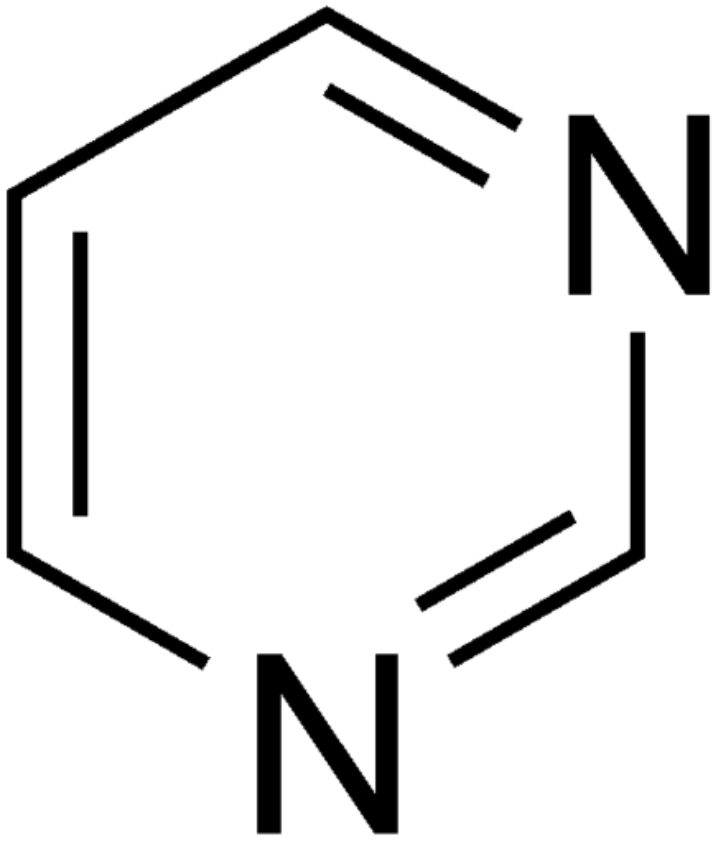
Base analogues

Nucleobase structure and nomenclature

Natural bases are divided into two classes depending on their structure: pyrimidine (an heterocyclic aromatic six-membered ring with nitrogen atoms in position 1 and 3) and purine (a pyrimidine (numeration inverted) fused with an imidazole ring, a five-membered ring with 2 nitrogen atoms separated by one carbon (meta), 7,9). Their main proprieties are base pairing, resulting form 2 or 3 hydrogen bonds between ketone (electron withdrawing group,*ei.* more negatively charged) and amino (electron withdrawing group,*ei.* more positively charged) functional groups, and base stacking, caused by the attraction of the delocalized π electron clouds of the aromatic ring structure.

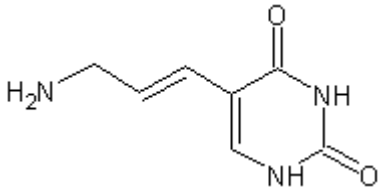


Purine



Pyrimidine

Fluorophores



Structure of aminoallyl-uridine

Commonly fluorophores (such as rhodamine or fluorescein) are linked to the ring linked to the sugar (in para) via a flexible arm, presumably extruding from the major groove of the helix. Due to low processivity of the nucleotides linked to bulky adducts such as florophores by taq polymerases, the sequence is typically copied using a nucleotide with an arm and later coupled with a reactive fluorophore (indirect labelling):

- amine reactive: Aminoallyl nucleotide contain a primary amine group on a linker that reacts with the amino-reactive dye such as a cyanine or Alexa Fluor dyes, which contain a reactive leaving group, such as a succinimidyl ester (NHS). (base-pairing amino groups are not affected).
- thiol reactive: thiol containing nucleotides reacts with the fluorophore linked to a reactive leaving group, such as a maleimide.
- biotin linked nucleotides rely on the same indirect labelling principle (+ fluorescent streptavidin) and are used in Affymetrix DNAchips.

Fluorophores find a variety of uses in medicine and biochemistry.

Fluorescent base analogues

The most commonly used and commercially available fluorescent base analogue, 2-aminopurine (2-AP), has a high-fluorescence quantum yield free in solution (0.68) that is considerably reduced (appr. 100 times but highly dependent on base sequence) when incorporated into nucleic acids. The emission sensitivity of 2-AP to immediate surroundings is shared by other promising and useful fluorescent base analogues like 3-MI, 6-MI, 6-MAP, pyrrolo-dC (also commercially available), modified and improved derivatives of pyrrolo-dC, furan-modified bases and many other ones. This sensitivity to the microenvironment have been utilized in studies of e.g. structure and dynamics within both DNA and RNA, dynamics and kinetics of DNA-protein interaction and electron transfer within DNA. A newly developed and very interesting group of fluorescent base analogues that has a fluorescence quantum yield that is nearly insensitive to their immediate surroundings is the tricyclic cytosine family. 1,3-Diaza-2-oxophenothiazine, tC, has a fluorescence quantum yield of approximately 0.2 both in single- and in double-strands irrespective of surrounding bases. Also the oxo-homologue of tC called tC^O (both commercially available), 1,3-diaza-2-oxophenoxazine, has a quantum yield of 0.2 in double-stranded systems. However, it is somewhat sensitive to surrounding bases in single-strands (quantum yields of 0.14–0.41). The high and stable quantum yields of these base analogues make them very bright, and, in combination with their good base analogue properties (leaves DNA structure and stability next to unperturbed), they are especially useful in fluorescence anisotropy and FRET measurements, areas where other fluorescent base analogues are less accurate. Also, in the same family of cytosine analogues, a FRET-acceptor base analogue, tC_{nitro}, has been developed. Together with tC^O as a FRET-donor this constitutes the first nucleic acid base analogue FRET-pair ever developed. The tC-family has, for example, been used in studies related to polymerase DNA-binding and DNA-polymerization mechanisms.

Natural non-canon bases

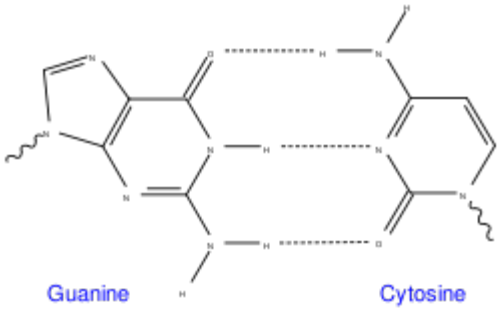
In a cell, there are several noncanon bases present: CpG islands in DNA (are often methylated), all eukaryotic mRNA (capped with a methyl-7-guanosine), and several bases of rRNAs (are methylated). Often, tRNAs are heavily modified postranscriptionally in order to improve their conformation or base pairing, in particular in/near the anticodon: inosine can base pair with C, U, and even with A, whereas thiouridine (with A) is more

specific than uracil (with a purine). Other common tRNA base modifications are pseudouridine (which gives its name to the TΨC loop), dihydrouridine (which does not stack as it is not aromatic), queuosine, wyosine, and so forth. Nevertheless these are all modifications to normal bases and are not placed by a polymerase.

Base-pairing

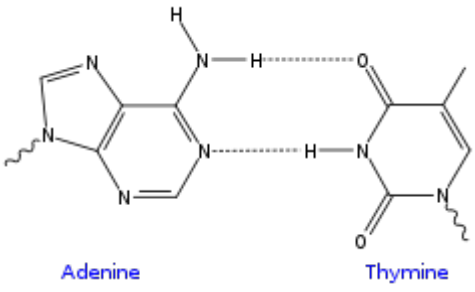
Canonical bases may have either a ketone or an amine group on the carbons surrounding the nitrogen atom furthest away from the glycosidic bond, which allows them to base pair (Watson-Crick base pairing) via hydrogen bonds (amine with ketone, purine with pyrimidine). Adenine and 2-aminoadenine have one/two amine group(s), whereas thymine has two ketone groups, and cytosine and guanine are mixed amine and ketone (inverted in respect to each other).

Natural basepairs



A GC basepair: purine ketone/amine forms three

intermolecular hydrogen bonds with pyrimidine amine/ketone



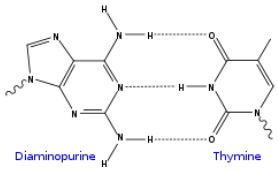
An AT basepair: purine amine/- forms two

intermolecular hydrogen bonds with pyrimidine ketone/ketone

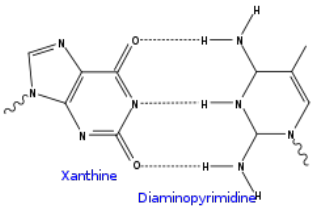
The precise reason why there are only four nucleotides is debated, but there are several unused possibilities. Furthermore, adenine is not the most stable choice for base pairing: in Cyanophage S-2L diaminopurine (DAP) is used instead of adenine (host evasion). Diaminopurine basepairs perfectly with thymine as it is identical to adenine but has an amine group at position 2 forming 3 intramolecular hydrogen bonds, eliminating the major difference between the two types of basepairs (Weak:A-T and Strong:C-G). This improved stability affects protein-binding ineractions that rely on those differences. Other combination include,

- isoguanine and isocytosine, which have their amine and ketone inverted compared to standard guanine and cytosine, (not used probably as tautomers are problematic for base pairing, but isoC and isoG can be amplified correctly with PCR even in the presence of the 4 canon bases)
- diaminopyrimidine and a xanthine, which bind like 2-aminoadenine and thymine but with inverted structures (not used as xanthine is a deamination product)

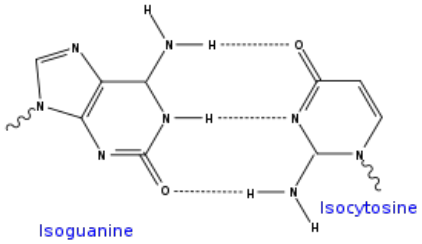
Unused basepair arrangements



A DAP-T base: purine amine/amine forms three intermolecular hydrogen bonds with pyrimidine ketone/ketone



An X-DAY base: purine ketone/ketone forms three intermolecular hydrogen bonds with pyrimidine amine/amine



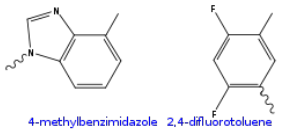
A iG-iC base: purine amine/ketone forms three intermolecular hydrogen bonds with pyrimidine ketone/amine

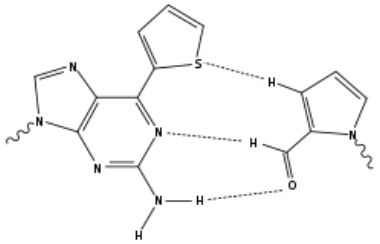
However, correct DNA structure can form even when the bases are not paired via hydrogen bonding; that is, the bases pair thanks to hydrophobicity, as studies have shown using DNA isosteres (analogues with same number of atoms), such as the thymine analogue 2,4-difluorotoluene (F) or the adenine analogue 4-methylbenzimidazole (Z). An alternative hydrophobic pair could be isoquinoline, and the pyrrolo[2,3-b]pyridine

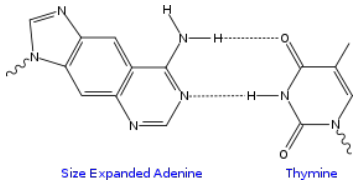
Other noteworthy basepairs:

- Several fluorescent bases have also been made, such as the 2-amino-6-(2-thienyl)purine and pyrrole-2-carbaldehyde base pair.
- Metal coordinated bases, such as two 2,6-bis(ethylthiomethyl)pyridine (SPy) with a silver ion or pyridine-2,6-dicarboxamide (Dipam) and a mondentate pyridine (Py) with a copper ion.
- Universal bases may pair indiscriminately with any other base, but, in general, lower the melting temperature of the sequence considerably; examples include 2'-deoxyinosine (hypoxanthine deoxynucleotide) derivatives, nitroazole analogues, and hydrophobic aromatic non-hydrogen-bonding bases (strong stacking effects). These are used as proof of concept and, in general, are not utilised in degenerate primers (which are a mixture of primers).
- The numbers of possible base pairs is doubled when xDNA is considered. xDNA contains expanded bases, in which a benzene ring has been added, which may pair with canon bases, resulting in four possible base-pairs (8 bases:xA-T,xT-A,xC-G,xG-C, 16 bases if the unused arrangements are used). Another form of benzene added bases is yDNA, in which the base is widened by the benzene.

Novel basepairs with special proprieties







A F-Z base: methylbenzimidazole does not form intermolecular hydrogen bonds with toluene F/F

An S-Pa base: purine thienyl/amine forms three intermolecular hydrogen bonds with pyrrole -/carbaldehyde

An xA-T base: same bonding as A-T

Metal Base Pairs

In metal base-pairing, the Watson-Crick hydrogen bonds are replaced by the interaction between a metal ion with nucleosides acting as ligands. The possible geometries of the metal that would allow for duplex formation with two bidentate nucleosides around a central metal atom are: tetrahedral, dodecahedral, and square planar. Metal-complexing with DNA can occur by the formation of non-canonical base pairs from natural nucleobases with participation by metal ions and also by the exchanging the hydrogen atoms that are part of the Watson-Crick base pairing by metal ions. Introduction of metal ions into a DNA duplex has shown to have potential magnetic, conducting properties, as well as increased stability.

Metal complexing has been shown to occur between natural nucleobases. A well-documented example is the formation of T-Hg-T, which involves two deportonated thymine nucleobases that are brought together by Hg^{2+} and forms a connected metal-base pair. This motif does not accommodate stacked Hg^{2+} in a duplex due to an intrastrand hairpin formation process that is favored over duplex formation. Two thymines across from each other in a duplex do not form a Watson-Crick base pair in a duplex; this is an example where a Watson-Crick basepair mismatch is stabilized by the formation of the metal-base pair. Another example of a metal complexing to natural nucleobases is the formation of A-Zn-T and G-Zn-C at high pH; Co^{+2} and Ni^{+2} also form these complexes. These are Watson-Crick base pairs where the divalent cation in coordinated to the nucleobases. The exact binding is debated.

A large variety of artificial nucleobases have been developed for use as metal base pairs. These modified nucleobases exhibit tunable electronic properties, sizes, and binding affinities that can be optimized for a specific metal. For, example a nucleoside modified with a pyridine-2,6-dicarboxylate has shown to bind tightly to Cu^{2+}, whereas other divalent ions are only loosely bound. The tridentate character contributes to this

selectivity. The fourth coordination site on the copper is saturated by an oppositely arranged pyridine nucleobase. The asymmetric metal base pairing system is orthogonal to the Watson-Crick base pairs. Another example of an artificial nucleobase is that with hydroxypyridone nucleobases, which are able to bind Cu^{2+} inside the DNA duplex. Five consecutive copper-hydroxypyridone base pairs were incorporated into a double strand, which were flanked by only one natural nucleobase on both ends. EPR data showed that the distance between copper centers was estimated to be 3.7 ± 0.1 Å, while a natural B-type DNA duplex is only slightly larger (3.4 Å). The appeal for stacking metal ions inside a DNA duplex is the hope to obtain nanoscopic self-assembling metal wires, though this has not been realized yet.

Orthogonal system

It has been proposed and studied both theoretically and experimentally the possibility of implementing an orthogonal system inside cells independent of the cellular genetic material in order to make a completely safe system, with the possible increase in encoding potentials Several groups have focused on different aspects:

- novel backbones and base pairs as discussed above
- XNA replication/transcription polymerases starting generally from T7 RNA polymerase
- rybozomes (16S sequences with altered anti Shine-Dalgarno sequence allowing the translation of only orthogonal mRNA with a matching altered Shine-Dalgarno sequence)
- novel tRNA encoding non-natural aminoacids.

Chapter- 6

Aptamer

Aptamers are oligonucleic acid or peptide molecules that bind to a specific target molecule. Aptamers are usually created by selecting them from a large random sequence pool, but natural aptamers also exist in riboswitches. Aptamers can be used for both basic research and clinical purposes as macromolecular drugs. Aptamers can be combined with ribozymes to self-cleave in the presence of their target molecule. These compound molecules have additional research, industrial and clinical applications.

More specifically, aptamers can be classified as:

- DNA or RNA aptamers. They consist of (usually short) strands of oligonucleotides.
- Peptide aptamers. They consist of a short variable peptide domain, attached at both ends to a protein scaffold.

Nucleic Acid aptamers

Nucleic acid aptamers are nucleic acid species that have been engineered through repeated rounds of *in vitro selection* or equivalently, SELEX (systematic evolution of ligands by exponential enrichment) to bind to various molecular targets such as small molecules, proteins, nucleic acids, and even cells, tissues and organisms. Aptamers are useful in biotechnological and therapeutic applications as they offer molecular recognition properties that rival that of the commonly used biomolecule, antibodies. In addition to their discriminate recognition, aptamers offer advantages over antibodies as they can be engineered completely in a test tube, are readily produced by chemical synthesis, possess desirable storage properties, and elicit little or no immunogenicity in therapeutic applications.

In 1990, two labs independently developed the technique of selection: the Gold lab, using the term SELEX for their process of selecting RNA ligands against T4 DNA polymerase; and the Szostak lab, coining the term *in vitro selection*, selecting RNA ligands against various organic dyes. The Szostak lab also coined the term aptamer (from the Latin, *aptus*, meaning ‘to fit’) for these nucleic acid-based ligands. Two years later, the Szostak

lab and Gilead Sciences, independent of one another, used *in vitro selection* schemes to evolve single stranded DNA ligands for organic dyes and human coagulant, thrombin, respectively. There does not appear to be any systematic differences between RNA and DNA aptamers, save the greater intrinsic chemical stability of DNA.

Interestingly enough, the notion of selection *in vitro* was actually preceded twenty-plus years prior when Sol Spiegelman used a Qbeta replication system as a way to evolve a self-replicating molecule. In addition, a year before the publishing of *in vitro selection* and SELEX, Gerald Joyce used a system that he termed 'directed evolution' to alter the cleavage activity of a ribozyme.

Since the discovery of aptamers, many researchers have used aptamer selection as a means for application and discovery. In 2001, the process of *in vitro selection* was automated by the Ellington lab at the University of Texas at Austin, and at SomaLogic, Inc (Boulder, CO), reducing the duration of a selection experiment from six weeks to three days.

While the process of artificial engineering of nucleic acid ligands is highly interesting to biology and biotechnology, the notion of aptamers in the natural world had yet to be uncovered until 2002 when two groups led by Ronald Breaker and Evgeny Nudler discovered a nucleic acid-based genetic regulatory element called a riboswitch that possesses similar molecular recognition properties to the artificially made aptamers. In addition to the discovery of a new mode of genetic regulation, this adds further credence to the notion of an 'RNA World,' a postulated stage in time in the origins of life on Earth.

Lately, a concept of smart aptamers, and smart ligands in general, has been introduced. It describes aptamers that are selected with pre-defined equilibrium (K_d), rate (k_{off}, k_{on}) constants and thermodynamic (ΔH, ΔS) parameters of aptamer-target interaction. Kinetic capillary electrophoresis is the technology used for the selection of smart aptamers. It obtains aptamers in a few rounds of selection.

Recent developments in aptamer-based therapeutics have been rewarded in the form of the first aptamer-based drug approved by the U.S. Food and Drug Administration (FDA) in treatment for age-related macular degeneration (AMD), called Macugen offered by OSI Pharmaceuticals. In addition, Cambridge, MA - based Archemix is leading the development of aptamers as a new class of directed therapeutics for the prevention and treatment of chronic and acute diseases. ARC1779, its lead proprietary candidate, is a potent, selective, first-in-class antagonist of von Willebrand Factor (vWF). ARC1779 is being evaluated in patients diagnosed with acute coronary syndrome (ACS) who are undergoing percutaneous coronary intervention (PCI). Phase I testing for ARC1779 was initiated in December 2006, and a Phase 2 study in ACS is planned to begin by the end of 2007.

Non-modified aptamers are cleared rapidly from the bloodstream, with a half-life of minutes to hours, mainly due to nuclease degradation and clearance from the body by the kidneys, a result of the aptamer's inherently low molecular weight. Unmodified aptamer

applications currently focus on treating transient conditions such as blood clotting, or treating organs such as the eye where local delivery is possible. This rapid clearance can be an advantage in applications such as *in vivo* diagnostic imaging. An example is a tenascin-binding aptamer under development by Schering AG for cancer imaging. Several modifications, such as 2'-fluorine-substituted pyrimidines, polyethylene glycol (PEG) linkage, etc. (both of which are used in Macugen, an FDA-approved aptamer) are available to scientists with which to increase the serum half-life of aptamers easily to the day or even week time scale.

In addition to the development of aptamer-based therapeutics, many researchers such as the Ellington lab and the Boulder, CO-based SomaLogic have been developing diagnostic techniques for whole cell protein profiling called proteomics, and medical diagnostics for the distinction of disease versus healthy states.

As a resource for all *in vitro selection* and SELEX experiments, the Ellington lab has developed the Aptamer Database cataloging all published experiments.

Peptide aptamers

Peptide aptamers are proteins that are designed to interfere with other protein interactions inside cells. They consist of a variable peptide loop attached at both ends to a protein scaffold. This double structural constraint greatly increases the binding affinity of the peptide aptamer to levels comparable to an antibody's (nanomolar range).

The variable loop length is typically composed of ten to twenty amino acids, and the scaffold may be any protein which has good solubility and compacity properties. Currently, the bacterial protein Thioredoxin-A is the most used scaffold protein, the variable loop being inserted within the reducing active site, which is a -Cys-Gly-Pro-Cys- loop in the wild protein, the two Cysteines lateral chains being able to form a disulfide bridge.

Peptide aptamer selection can be made using different systems, but the most used is currently the yeast two-hybrid system.

Selection of Ligand Regulated Peptide Aptamers (LiRPAs) has been demonstrated. By displaying 7 amino acid peptides from a novel scaffold protein based on the trimeric FKBP-rapamycin-FRB structure, interaction between the randomized peptide and target molecule can be controlled by the small molecule Rapamycin or non-immunosuppressive analogs.

AptaBiD

AptaBiD or Aptamer-Facilitated Biomarker Discovery is a technology for biomarker discovery. AptaBiD is based on multi-round generation of an aptamer or a pool of aptamers for differential molecular targets on the cells which facilitates exponential detection of biomarkers. It involves three major stages: (i) differential multi-round

selection of aptamers for biomarker of target cells; (ii) aptamer-based isolation of biomarkers from target cells; and (iii) mass spectrometry identification of biomarkers. The important feature of the AptaBiD technology is that it produces synthetic affinity probes (aptamers) simultaneously with biomarker discovery. In AptaBiD, aptamers are developed for cell surface biomarkers in their native state and conformation. In addition to facilitating biomarker identification, such aptamers can be directly used for cell isolation, cell visualization, and tracking cells *in vivo*. They can also be used to modulate activities of cell receptors and deliver different agents (e.g., siRNA and drugs) into the cells.

Chapter- 7

Small Nucleolar RNA

Small nucleolar RNAs (**snoRNAs**) are a class of small RNA molecules that primarily guide chemical modifications of other RNAs, mainly ribosomal RNAs, transfer RNAs and small nuclear RNAs. There are two main classes of snoRNA, the C/D box snoRNAs which are associated with methylation, and the H/ACA box snoRNAs which are associated with pseudouridylation. snoRNAs are commonly referred to as guide RNAs but should not be confused with the guide RNAs that direct RNA editing in trypanosomes.

snoRNA guided modifications

After transcription, nascent rRNA molecules (termed pre-rRNA) are required to undergo a series of processing steps in order to generate the mature rRNA molecule. Prior to cleavage by exo- and endonucleases the pre-rRNA undergoes a complex pattern of nucleoside modifications. These include methylations and pseudouridylations, guided by snoRNAs.

- Methylation is the attachment or substitution of a methyl group onto various substrates. The rRNA of humans contain approximately 115 methyl group modifications. The majority of these are 2'O-ribose-methylations (where the methyl group is attached to the ribose group).

- Pseudouridylation is the conversion (isomerisation) of the nucleoside uridine to a different isomeric form pseudouridine(Ψ). Mature human rRNAs contain approximately 95 Ψ modifications.

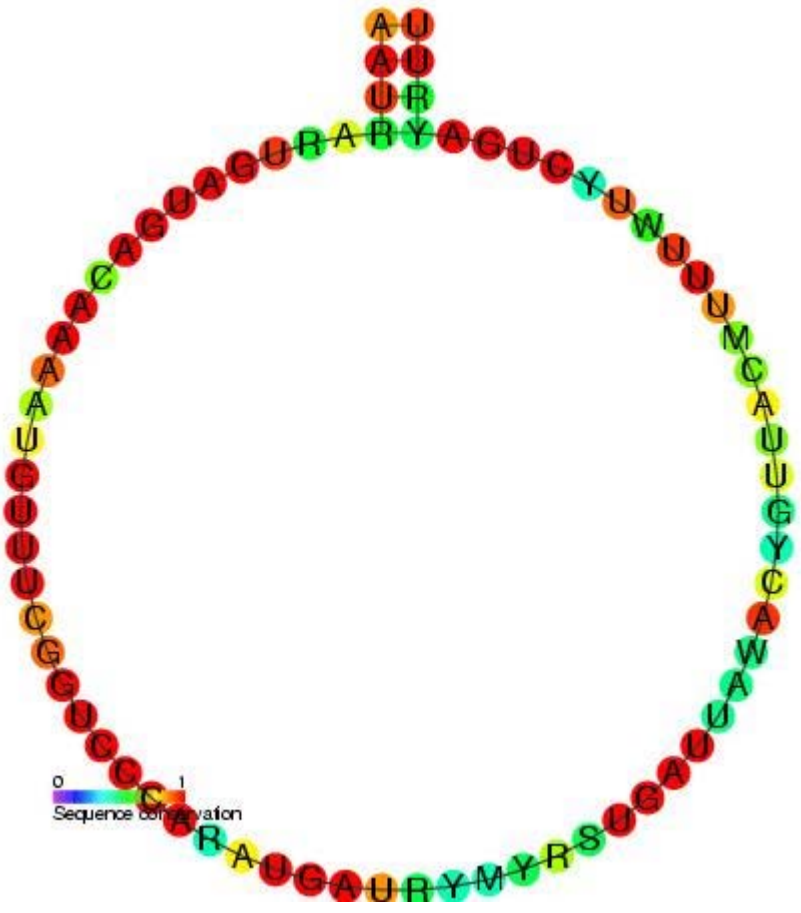
Each snoRNA molecule acts as a guide for only one (or two) individual modifications in a target RNA. In order to carry out modification, each snoRNA associates with at least four protein molecules in an RNA/protein complex referred to as a small nucleolar ribonucleoprotein (snoRNP). The proteins associated with each RNA depend on the type of snoRNA molecule. The snoRNA molecule contains an antisense element (a stretch of 10-20 nucleotides) which are base complementary to the sequence surrounding the base (nucleotide) targeted for modification in the pre-RNA molecule. This enables the

snoRNP to recognise and bind to the target RNA. Once the snoRNP has bound to the target site the associated proteins are in the correct physical location to catalyse the chemical modification of the target base.

snoRNA guide families

The two different types of rRNA modification (methylation and pseudouridylation) are directed by two different families of snoRNPs. These families of snoRNAs are referred to as antisense C/D box and H/ACA box snoRNAs based on the presence of conserved sequence motifs in the snoRNA. There are exceptions but as a general rule C/D box members guide methylation and H/ACA members guide pseudouridylation. The members of each family may vary in biogenesis, structure and function but each family is classified by the following generalised characteristics. SnoRNAs are classified under small nuclear RNA in MeSH.

C/D box

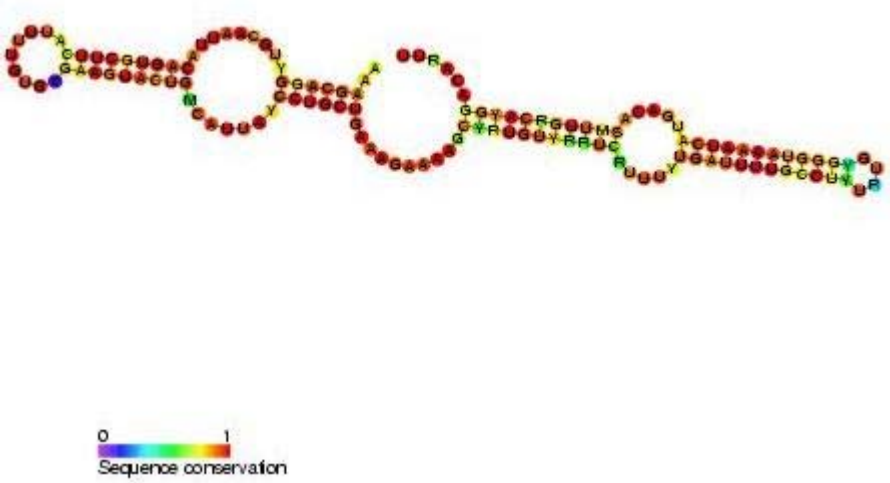


Example of a C/D box snoRNA secondary structure taken from the Rfam database. This example is SNORD73 (RF00071).

C/D box snoRNAs contain two short conserved sequence motifs, C (UGAUGA) and D (CUGA) located near the 5' and 3' ends of the snoRNA respectively. Short regions (~ 5 nucleotides) located upstream of the C box and downstream of the D box are usually base complementary and form a stem-box structure which brings the C and D box motifs into close proximity. This stem-box structure has been shown to be essential for correct snoRNA synthesis and nucleolar localization. Many C/D box snoRNA also contain an additional less well conserved copy of the C and D motifs (referred to as C' and D') located in the central portion of the snoRNA molecule. A conserved region of 10-21 nucleotides upstream of the D box is complementary to the methylation site of the target RNA and enables the snoRNA to form an RNA duplex with the RNA. The nucleotide to be modified in the target RNA is usually located at the 5th position upstream from the D box (or D' box). C/D box snoRNAs associate with four evolutionary conserved and essential proteins (Fibrillarin (Nop1p), Nop56p, Nop58p, and Snu13 (in eukaryotes; its archaeal homolog is L7Ae)) which make up the core C/D box snoRNP.

There exists a eukaryotic C/D box snoRNA (snoRNA U3) that has not been shown to guide 2'-*O*-methylation. Instead, it functions in rRNA processing by directing pre-rRNA cleavage.

H/ACA box



Example of a H/ACA box snoRNA secondary structure taken from the Rfam database. This example is SNORA69 (RF00265).

H/ACA box snoRNAs have a common secondary structure consisting of a two hairpins and two single stranded regions termed a hairpin-hinge-hairpin-tail structure. H/ACA snoRNAs also contain conserved sequence motifs known as H box (consensus ANANNA) and the ACA box (ACA). Both motifs are usually located in the single stranded regions of the secondary structure. The H motif is located in the hinge and the ACA motif is located in the tail region, 3 nucleotides from the 3' end of the sequence. The hairpin regions contain internal bulges known as recognition loops in which the antisense guide sequences (bases complementary to the target sequence) are located. This recognition sequence is bipartite (constructed from the two different arms of the loop

region) and forms complex pseudo-knots with the target RNA. H/ACA box snoRNAs associate with four evolutionary conserved and essential proteins (dyskerin (Cbf5p), Gar1p, Nhp2p and Nop10p) which make up the core of the H/ACA box snoRNP. However, in lower eukaryotic cells such as trypanosomes, similar RNAs exist in the form of single hairpin structure and an AGA box instead of ACA box at the 3' end of the RNA.

The RNA component of human telomerase (hTERC) contains an H/ACA domain for pre-RNP formation and nucleolar localization of the telomerase RNP itself. Importantly, the H/ACA snoRNP has been implicated in the rare genetic disease dyskeratosis congenita (DKC) due to its affiliation with human telomerase. Mutations in the protein component of the H/ACA snoRNP result in a reduction in physiological TERC levels. This has been strongly correlated with the pathology behind DKC which seems to be primarily a disease of poor telomere maintenance.

Composite H/ACA and C/D box

An unusual guide snoRNA U85 was identified that functions in both 2'-O-ribose methylation and pseudouridylation of small nuclear RNA (snRNA) U5. This composite snoRNA contains both C/D and H/ACA box domains and associates with the proteins specific to each class of snoRNA (fibrillaring and Gar1p respectively. More composite snoRNAs have now been characterised.

These composite snoRNAs have been found to accumulate in a subnuclear organelle called the Cajal body and are referred to as Cajal body specific RNAs. This is in contrast to the majority of C/D box or H/ACA box snoRNAs which localise to the nucleolus. These Cajal body specific RNAs and are proposed to be involved in the modification of RNA polymerase II transcribed spliceosomal RNAs U1, U2, U4, U5 and U12. Not all snoRNAs that have been localised to Cajal bodies are composite C/D and H/ACA box snoRNAs.

Orphan snoRNAs

The targets for newly identified snoRNAs are predicted on the basis of sequence complementarity between putative target RNAs and the antisense elements or recognition loops in the snoRNA sequence. However, there are an increasing number of 'orphan' guides without any known RNA targets, which suggests that there might be more proteins or transcripts involved in rRNA than previously and/or that some snoRNAs have different functions not concerning rRNA. There is evidence that some of these orphan snoRNAs regulate alternatively spliced transcripts. For example, it appears that the C/D box snoRNA SNORD115 regulates the alternative splicing of the serotonin 2C receptor mRNA via a conserved region of complementarity. Another C/D box snoRNA, SNORD116, that resides in the same cluster as SNORD115 has been predicted to have 23 possible targets within protein coding genes using a bioinformatic approach. Of these a large fraction were found to be alternatively spliced, suggesting a role of SNORD116 in the regulation of alternative splicing.

Target modifications

The precise effect of the methylation and pseudouridylation modifications on the function of the mature RNAs is not yet known. The modifications do not appear to be essential but are known to subtly enhance the RNA folding and interaction with ribosomal proteins. In support of their importance, target site modifications are exclusively located within conserved and functionally important domains of the mature RNA and are commonly conserved amongst distant eukaryotes.

1. 2'-O-methylated ribose causes an increase in the 3'-endo conformation
2. Pseudouridine (psi/Ψ) adds another option for H-bonding.
3. Heavily methylated RNA is protected from hydrolysis. rRNA acts as a ribozyme by catalyzing its own hydrolysis and splicing.

Genomic organisation

The majority of vertebrate snoRNA genes are encoded in the introns of proteins involved in ribosome synthesis or translation, and are synthesized by RNA polymerase II, but snoRNAs can also be transcribed from their own promoters by RNA polymerase II or III.

Imprinted loci

In the human genome there are at least two examples where C/D box snoRNAs are found in tandem repeats within imprinted loci. These two loci (14q32 on chromosome 14 and 15q11q13 on chromosome 15) have been extensively characterised and in both regions multiple snoRNAs have been found located within introns in clusters of closely related copies.

In 15q11q13, five different snoRNAs have been identified (SNORD64, SNORD107, SNORD108, SNORD109 (two copies), SNORD116 (29 copies) and SNORD115 (48 copies). Loss of the 29 copies of SNORD116 from this region has been identified as a cause of Prader-Willi syndrome whereas gain of additional copies of SNORD115 has been linked to autism.

Region 14q32 contains repeats of two snoRNAs SNORD113 (9 copies) and SNORD114 (31 copies) within the introns of a tissue-specific ncRNA transcript (MEG8). The 14q32 domain has been shown to share common genomic features with the imprinted 15q11-q13 loci and a possible role for tandem repeats of C/D box snoRNAs in the evolution or mechanism of imprinted loci has been suggested.

Other functions of snoRNA

snoRNAs can function as miRNAs. It has been shown that human ACA45 is a *bona fide* snoRNA that can be processed into a 21 nucleotides long mature miRNA by the RNAse III family endoribonuclease dicer. This snoRNA product has previously been identified as mmu-miR-1839 and was shown to be processed independent of the other miRNA

generating endoribonuclease drosha. Bioinformatical analyses have revealed that putatively snoRNA-derived, miRNA-like fragments occur in different organisms.

Recently, it has been found that snoRNAs can have functions not related to rRNA. One such function is the regulation of alternative splicing of the *trans* gene transcript, which is done by the snoRNA HBII-52, which is also known as SNORD115.

Recent discovery also show, the existence of snoRNA, microRNA, piRNA characteristics in a novel non-coding RNA: x-ncRNA and its biological implication in *Homo sapiens*.

Chapter- 8

Glycol Nucleic Acid and Locked Nucleic Acid

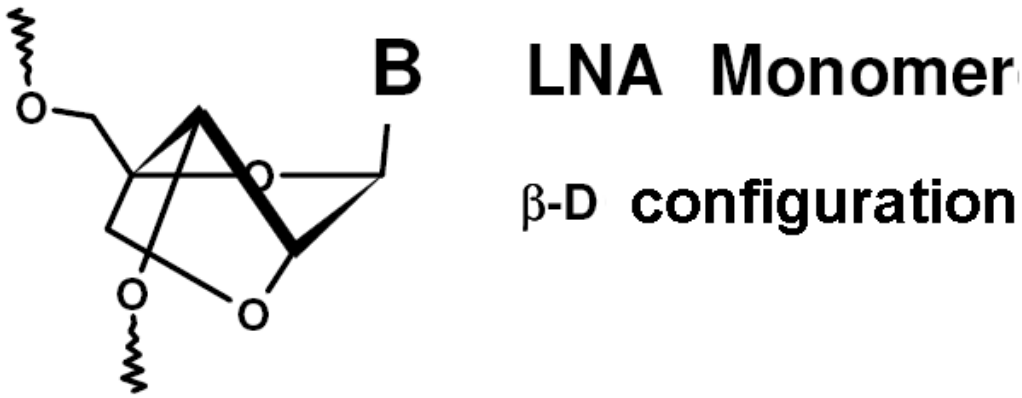
Glycol nucleic acid

Glycol nucleic acid (GNA) is a polymer similar to DNA or RNA but differing in the composition of its "backbone". GNA is not known to occur naturally.

The 2,3-dihydroxypropylnucleoside analogues were first prepared by Ueda et al. (1971). Soon thereafter it was shown that phosphate-linked oligomers of the analogues do in fact exhibit hypochromicity in the presence of RNA and DNA in solution (Seita et al. 1972). The preparation of the polymers was later described by Cook et al. (1995, 1999) and Acevedo and Andrews (1996). The GNA-GNA self-pairing described by Zhang and Meggers is however novel, and the specificity of interaction well-demonstrated.

DNA and RNA have a deoxyribose and ribose sugar backbone, respectively, whereas GNA's backbone is composed of repeating glycerol units linked by phosphodiester bonds. The glycerol molecule has just three carbon atoms and still shows Watson-Crick base pairing. Interestingly, the Watson-Crick base pairing is much more stable in GNA than its natural counterparts DNA and RNA as it requires a high temperature to melt a duplex of GNA. It is possibly the simplest of the nucleic acids, so making it a hypothetical precursor to RNA.

Locked nucleic acid



Structure of an LNA monomer

A **locked nucleic acid** (LNA), often referred to as inaccessible RNA, is a modified RNA nucleotide. The ribose moiety of an LNA nucleotide is modified with an extra bridge connecting the 2' oxygen and 4' carbon. The bridge "locks" the ribose in the 3'-*endo* (North) conformation, which is often found in the A-form duplexes. LNA nucleotides can be mixed with DNA or RNA residues in the oligonucleotide whenever desired. Such oligomers are commercially available. The locked ribose conformation enhances base stacking and backbone pre-organization. This significantly increases the hybridization properties (melting temperature) of oligonucleotides.

LNA was independently synthesized by the group of Jesper Wengel in 1998, soon after the first synthesis by the group of Takeshi Imanishi in 1997. The exclusive rights to the LNA technology were secured in 1997 by Exiqon A/S, a Danish biotech company.

LNA nucleotides are used to increase the sensitivity and specificity of expression in DNA microarrays, FISH probes, real-time PCR probes and other molecular biology techniques based on oligonucleotides. For the *in situ* detection of miRNA the use of LNA is currently (2005) the only efficient method. A triplet of LNA nucleotides surrounding a single-base mismatch site maximizes LNA probe specificity unless the probe contains the guanine base of G-T mismatch.

Using LNA based oligonucleotides therapeutically is an emerging field in biotechnology. The Danish pharmaceutical company Santaris Pharma a/s owns the sole rights to therapeutic uses of LNA technology, and is now developing a new, LNA based, hepatitis C drug called miravirsen, targeting miR-122, which is in Phase II clinical testing as of late 2010.

Benefits of the LNA technology
Some of the benefits of using LNA include:

- Ideal for the detection of short RNA and DNA targets
- Increases the thermal stability of duplexes
- Capable of single nucleotide discrimination
- Resistant to exo- and endonucleases resulting in high stability in vivo and in vitro applications
- Increased target specificity
- Facilitates Tm normalization
- Strand invasion properties enables detection of "hard to access" samples
- Compatible with standard enzymatic processes

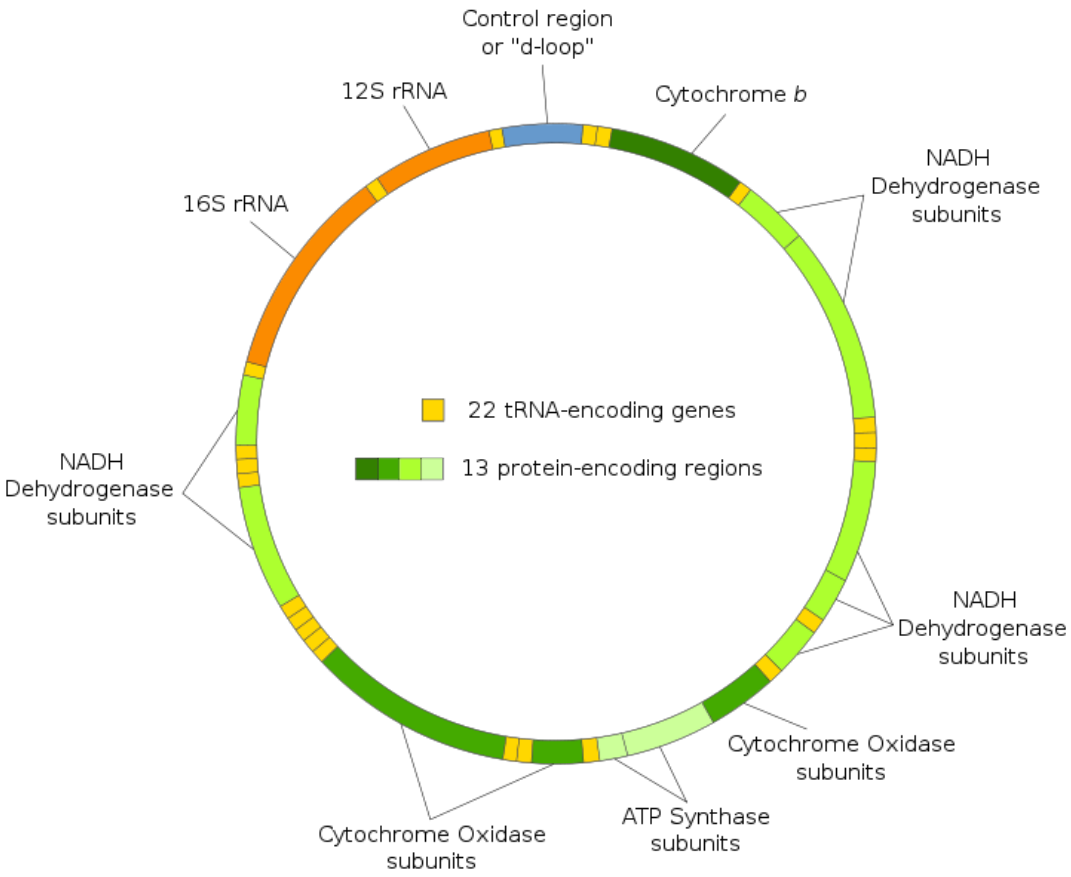
Applications of the LNA technology
Some proven applications of LNA include:

- Allele-specific PCR: allows for the design of shorter primers, without compromising binding specificity
- Microarray gene expression profiling: provides increased sensitivity and selectivity with smaller amounts of substrates
- Small RNA research
- SNP genotyping
- mRNA antisense oligonucleotides
- RNAi
- DNAzymes
- Fluorescence Polarization probes
- Molecular Beacons
- Gene repair/exon skipping
- Splice variant detection
- Comparative genome hybridization (GCH)

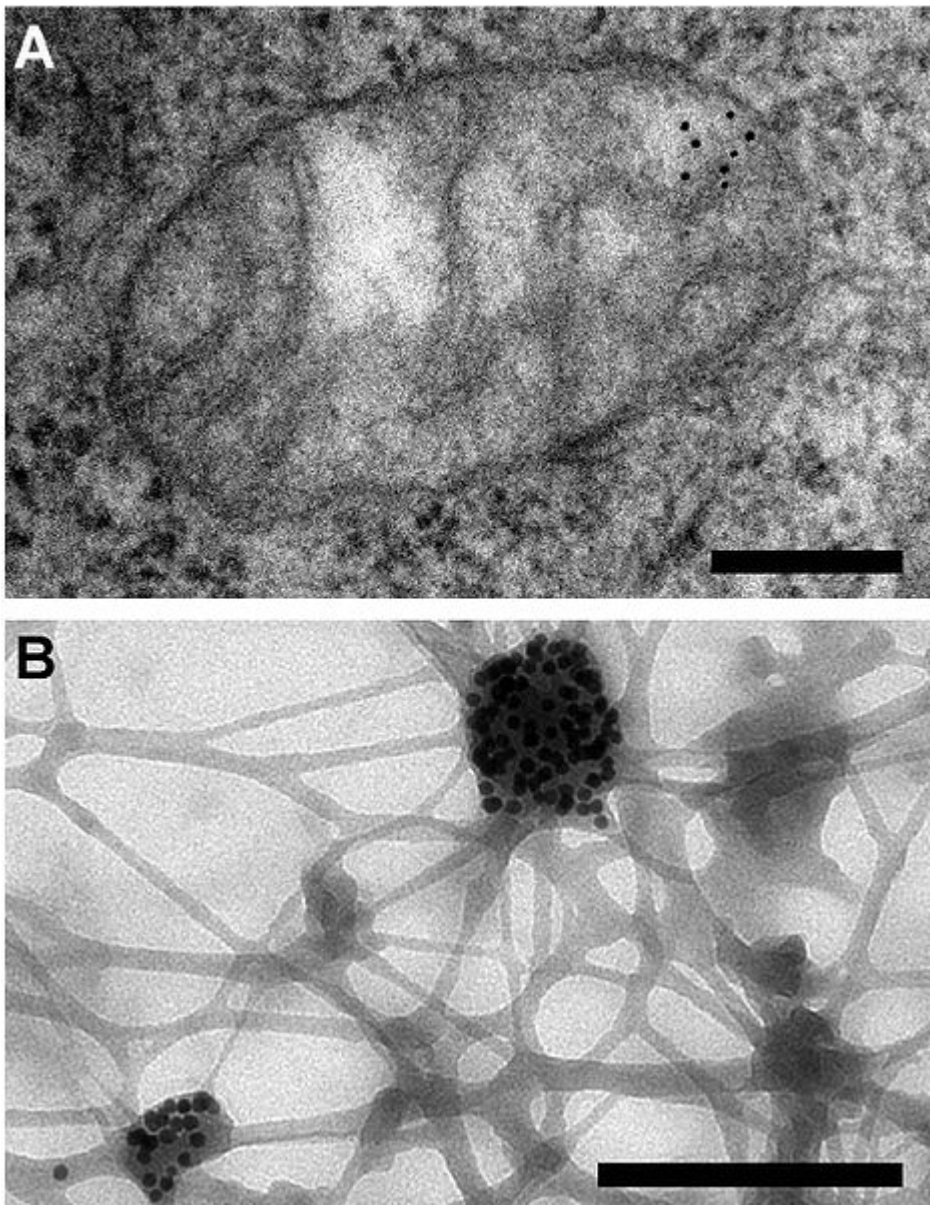
Other therapeutic and diagnostic applications of LNA technology are in development.

Chapter- 9

Mitochondrial DNA



Mitochondrial DNA



Electron microscopy reveals mitochondrial DNA in discrete foci. Bars: 200 nm. (A) Cytoplasmic section after immunogold labelling with anti-DNA; gold particles marking mtDNA are found near the mitochondrial membrane. (B) Whole mount view of cytoplasm after extraction with CSK buffer and immunogold labelling with anti-DNA; mtDNA (marked by gold particles) resists extraction.

Mitochondrial DNA (**mtDNA**) is the DNA located in organelles called mitochondria, structures within eukaryotic cells that convert the chemical energy from food into a form that cells can use, ATP. Most other DNA present in eukaryotic organisms is found in the cell nucleus.

Replication

mtDNA is replicated by the DNA polymerase gamma complex which is composed of a 140 kDa catalytic DNA polymerase encoded by the *POLG* gene and a 55 kDa accessory subunit encoded by the *POLG2* gene. During embryogenesis, replication of mtDNA is strictly down-regulated from the fertilized oocyte through the preimplantation embryo. At the blastocyst stage, the onset of mtDNA replication is specific to the cells of the trophectoderm. In contrast, the cells of the inner cell mass restrict mtDNA replication until they receive the signals to differentiate to specific cell types.

Origin

Nuclear and mitochondrial DNA are thought to be of separate evolutionary origin, with the mtDNA being derived from the circular genomes of the bacteria that were engulfed by the early ancestors of today's eukaryotic cells. This theory is called the endosymbiotic theory. Each mitochondrion is estimated to contain 2-10 mtDNA copies. In the cells of extant organisms, the vast majority of the proteins present in the mitochondria (numbering approximately 1500 different types in mammals) are coded for by nuclear DNA, but the genes for some of them, if not most, are thought to have originally been of bacterial origin, having since been transferred to the eukaryotic nucleus during evolution.

Mitochondrial inheritance

In most multicellular organisms, mtDNA is inherited from the mother (maternally inherited). Mechanisms for this include simple dilution (an egg contains 100,000 to 1,000,000 mtDNA molecules, whereas a sperm contains only 100 to 1000), degradation of sperm mtDNA in the fertilized egg, and, at least in a few organisms, failure of sperm mtDNA to enter the egg. Whatever the mechanism, this single parent (uniparental) pattern of mtDNA inheritance is found in most animals, most plants and in fungi as well.

Female inheritance

In sexual reproduction, mitochondria are normally inherited exclusively from the mother. The mitochondria in mammalian sperm are usually destroyed by the egg cell after fertilization. Also, most mitochondria are present at the base of the sperm's tail, which is used for propelling the sperm cells. Sometimes the tail is lost during fertilization. In 1999 it was reported that paternal sperm mitochondria (containing mtDNA) are marked with ubiquitin to select them for later destruction inside the embryo. Some *in vitro* fertilization techniques, particularly injecting a sperm into an oocyte, may interfere with this.

The fact that mitochondrial DNA is maternally inherited enables researchers to trace maternal lineage far back in time. (Y chromosomal DNA, paternally inherited, is used in an analogous way to trace the agnate lineage.) This is accomplished in humans by sequencing one or more of the hypervariable control regions (HVR1 or HVR2) of the mitochondrial DNA, as with a genealogical DNA test. HVR1 consists of about 440 base pairs. These 440 base pairs are then compared to the control regions of other individuals

(either specific people or subjects in a database) to determine maternal lineage. Most often, the comparison is made to the revised Cambridge Reference Sequence. Vilà *et al.* have published studies tracing the matrilineal descent of domestic dogs to wolves. The concept of the Mitochondrial Eve is based on the same type of analysis, attempting to discover the origin of humanity by tracking the lineage back in time.

Because mtDNA is not highly conserved and has a rapid mutation rate, it is useful for studying the evolutionary relationships - phylogeny - of organisms. Biologists can determine and then compare mtDNA sequences among different species and use the comparisons to build an evolutionary tree for the species examined.

Because mtDNA is transmitted from mother to child (both male and female), it can be a useful tool in genealogical research into a person's maternal line.

Male inheritance

It has been reported that mitochondria can occasionally be inherited from the father in some species such as mussels. Paternally inherited mitochondria have additionally been reported in some insects such as fruit flies, honeybees, and periodical cicadas.

Evidence supports rare instances of male mitochondrial inheritance in some mammals as well. Specifically, documented occurrences exist for mice, where the male-inherited mitochondria was subsequently rejected. It has also been found in sheep, and in cloned cattle. It has been found in a single case in a human male and was linked to infertility.

While many of these cases involve cloned embryos or subsequent rejection of the paternal mitochondria, others document *in vivo* inheritance and persistence under lab conditions.

Structure

In humans (and probably in metazoans in general), 100-10,000 separate copies of mtDNA are usually present per cell (egg and sperm cells are exceptions). In mammals, each double-stranded circular mtDNA molecule consists of 15,000-17,000 base pairs. The two strands of mtDNA are differentiated by their nucleotide content with the guanine rich strand referred to as the heavy strand, and the cytosine rich strand referred to as the light strand. The heavy strand encodes 28 genes, and the light strand encodes 9 genes for a total of 37 genes. Of the 37 genes, 13 are for proteins (polypeptides), 22 are for transfer RNA (tRNA) and two are for the small and large subunits of ribosomal RNA (rRNA). This pattern is also seen among most metazoans, although in some cases one or more of the 37 genes is absent and the mtDNA size range is greater. Even greater variation in mtDNA gene content and size exists among fungi and plants, although there appears to be a core subset of genes that are present in all eukaryotes (except for the few that have no mitochondria at all). Some plant species have enormous mtDNAs (as many as 2,500,000 base pairs per mtDNA molecule) but, surprisingly, even those huge mtDNAs

contain the same number and kinds of genes as related plants with much smaller mtDNAs.

Genes

Transport chain

The mitochondrial genome contains 13 protein-coding genes. Many of these genes encode the transport chain:

Category	Genes
NADH dehydrogenase (complex I)	MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-ND6
Coenzyme Q - cytochrome c reductase/Cytochrome b (complex III)	MT-CYB
cytochrome c oxidase (complex IV)	MT-CO1, MT-CO2, MT-CO3
ATP synthase	MT-ATP6, MT-ATP8

rRNA

Mitochondrial rRNA is encoded by MT-RNR1 (12S) and MT-RNR2 (16S).

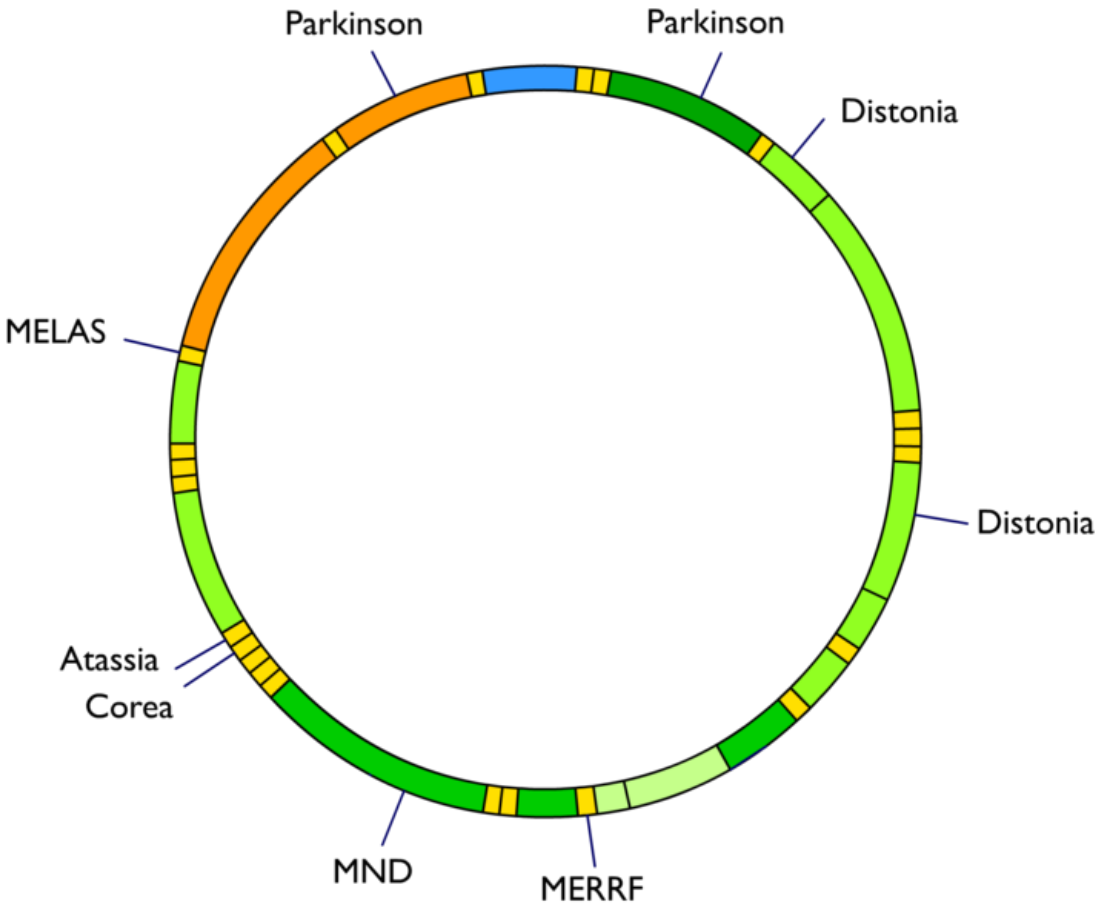
tRNA

The following genes encode tRNA:

Amino Acid	3-Letter	1-Letter	MT DNA
Alanine	Ala	A	MT-TA
Arginine	Arg	R	MT-TR
Asparagine	Asn	N	MT-TN
Aspartic acid	Asp	D	MT-TD
Cysteine	Cys	C	MT-TC
Glutamic acid	Glu	E	MT-TE
Glutamine	Gln	Q	MT-TQ
Glycine	Gly	G	MT-TG
Histidine	His	H	MT-TH
Isoleucine	Ile	I	MT-TI
Leucine	Leu	L	MT-TL1, MT-TL2
Lysine	Lys	K	MT-TK
Methionine	Met	M	MT-TM

Phenylalanine	Phe	F	MT-TF
Proline	Pro	P	MT-TP
Serine	Ser	S	MT-TS1, MT-TS2
Threonine	Thr	T	MT-TT
Tryptophan	Trp	W	MT-TW
Tyrosine	Tyr	Y	MT-TY
Valine	Val	V	MT-TV

Mutations



The involvement of mitochondrial DNA in several human diseases

Susceptibility

mtDNA is particularly susceptible to reactive oxygen species generated by the respiratory chain due to its close proximity. Though mtDNA is packaged by proteins and harbors significant DNA repair capacity, these protective functions are less robust than those operating on nuclear DNA and therefore thought to contribute to enhanced susceptibility of mtDNA to oxidative damage.

Genetic illness

Mutations of mitochondrial DNA can lead to a number of illnesses including exercise intolerance and Kearns-Sayre syndrome (KSS), which causes a person to lose full function of heart, eye, and muscle movements. Some evidence suggests that they might be major contributors to the aging process and age-associated pathologies.

Use in identification

In humans, mitochondrial DNA spans 16,569 DNA building blocks (base pairs), representing a fraction of the total DNA in cells. Unlike nuclear DNA, which is inherited from both parents and in which genes are rearranged in the process of recombination, there is usually no change in mtDNA from parent to offspring. Although mtDNA also recombines, it does so with copies of itself within the same mitochondrion. Because of this and because the mutation rate of animal mtDNA is higher than that of nuclear DNA, mtDNA is a powerful tool for tracking ancestry through females (matrilineage) and has been used in this role to track the ancestry of many species back hundreds of generations.

Human mtDNA can also be used to help identify individuals. Forensic laboratories occasionally use mtDNA comparison to identify human remains, and especially to identify older unidentified skeletal remains. Although unlike nuclear DNA, mtDNA is not specific to one individual, it can be used in combination with other evidence (anthropological evidence, circumstantial evidence, and the like) to establish identification. mtDNA is also used to exclude possible matches between missing persons and unidentified remains. Many researchers believe that mtDNA is better suited to identification of older skeletal remains than nuclear DNA because the greater number of copies of mtDNA per cell increases the chance of obtaining a useful sample, and because a match with a living relative is possible even if numerous maternal generations separate the two. American outlaw Jesse James's remains were identified using a comparison between mtDNA extracted from his remains and the mtDNA of the son of the female-line great-granddaughter of his sister. Similarly, the remains of Alexandra Feodorovna (Alix of Hesse), last Empress of Russia, and her children were identified by comparison of their mitochondrial DNA with that of Prince Philip, Duke of Edinburgh, whose maternal grandmother was Alexandra's sister Victoria of Hesse. Similarly to identify Emperor Nicholas II remains his mitochondrial DNA was compared with that of James Carnegie, 3rd Duke of Fife, whose maternal great-grandmother Alexandra of Denmark (Queen Alexandra) was sister of Nicholas II mother Dagmar of Denmark (Empress Maria Feodorovna).

The low effective population size and rapid mutation rate (in animals) makes mtDNA useful for assessing genetic relationships of individuals or groups within a species and also for identifying and quantifying the phylogeny (evolutionary relationships) among different species, provided they are not too distantly related. To do this, biologists determine and then compare the mtDNA sequences from different individuals or species. Data from the comparisons is used to construct a network of relationships among the sequences, which provides an estimate of the relationships among the individuals or

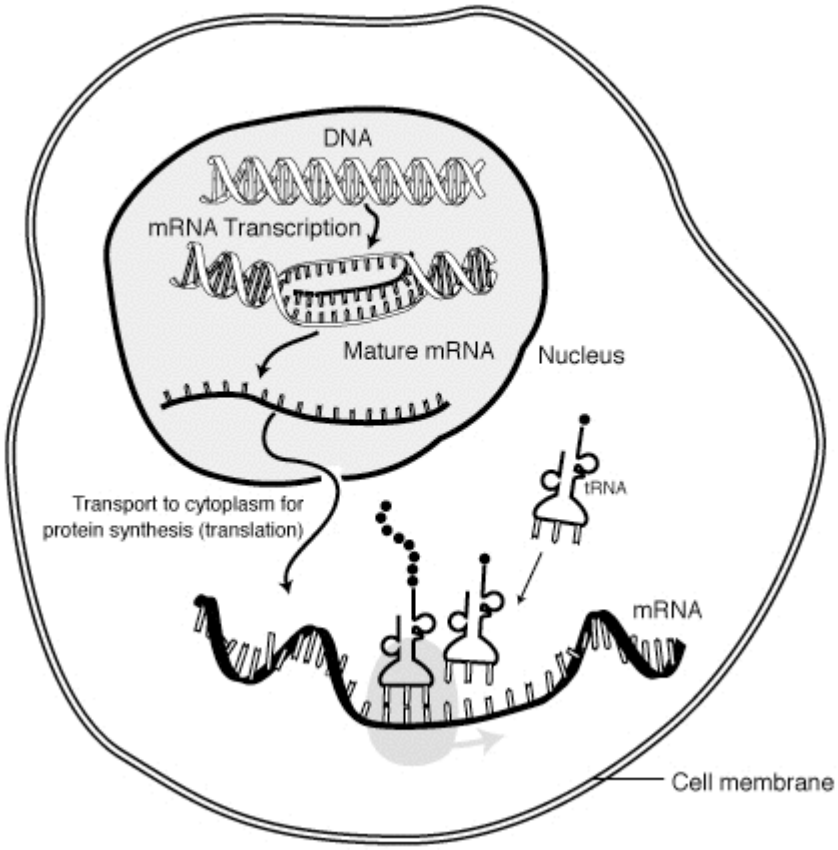
species from which the mtDNAs were taken. This approach has limits that are imposed by the rate of mtDNA sequence change. In animals, the high mutation rate makes mtDNA most useful for comparisons of individuals within species and for comparisons of species that are closely or moderately-closely related, among which the number of sequence differences can be easily counted. As the species become more distantly related, the number of sequence differences becomes very large; changes begin to accumulate on changes until an accurate count becomes impossible.

History

Mitochondrial DNA was discovered in the 1960s by Margit M. K. Nass and Sylvan Nass by electron microscopy as DNase-sensitive thread inside mitochondria, and by Ellen Haslbrunner, Hans Tuppy and Gottfried Schatz by biochemical assays on highly purified mitochondrial fractions.

Chapter- 10

Messenger RNA



The "life cycle" of an **mRNA** in a eukaryotic cell. RNA is transcribed in the nucleus; processed, it is transported to the cytoplasm and translated by the ribosome. At the end of its life, the mRNA is degraded.

Messenger RNA (**mRNA**) is a molecule of RNA encoding a chemical "blueprint" for a protein product. mRNA is transcribed from a DNA template, and carries coding information to the sites of protein synthesis: the ribosomes. Here, the nucleic acid polymer is translated into a polymer of amino acids: a protein. In mRNA as in DNA, genetic information is encoded in the sequence of nucleotides arranged into codons consisting of three bases each. Each codon encodes for a specific amino acid, except the

stop codons that terminate protein synthesis. This process requires two other types of RNA: transfer RNA (tRNA) mediates recognition of the codon and provides the corresponding amino acid, while ribosomal RNA (rRNA) is the central component of the ribosome's protein manufacturing machinery.

Synthesis, processing, and function

The brief existence of an mRNA molecule begins with transcription and ultimately ends in degradation. During its life, an mRNA molecule may also be processed, edited, and transported prior to translation. Eukaryotic mRNA molecules often require extensive processing and transport, while prokaryotic molecules do not.

Transcription

During transcription, RNA polymerase makes a copy of a gene from the DNA to mRNA as needed. This process is similar in eukaryotes and prokaryotes. One notable difference, however, is that prokaryotic RNA polymerase associates with mRNA processing enzymes during transcription so that processing can proceed quickly after the start of transcription. The short-lived, unprocessed or partially processed, product is termed *pre-mRNA*; once completely processed, it is termed *mature mRNA*.

Eukaryotic pre-mRNA processing

Processing of mRNA differs greatly among eukaryotes, bacteria and archea. Non-eukaryotic mRNA is essentially mature upon transcription and requires no processing, except in rare cases. Eukaryotic pre-mRNA, however, requires extensive processing.

5' cap addition

A *5′ cap* (also termed an RNA cap, an RNA 7-methylguanosine cap or an RNA m^7G cap) is a modified guanine nucleotide that has been added to the "front" or 5' end of a eukaryotic messenger RNA shortly after the start of transcription. The 5' cap consists of a terminal 7-methylguanosine residue which is linked through a 5'-5'-triphosphate bond to the first transcribed nucleotide. Its presence is critical for recognition by the ribosome and protection from RNases.

Cap addition is coupled to transcription, and occurs co-transcriptionally, such that each influences the other. Shortly after the start of transcription, the 5' end of the mRNA being synthesized is bound by a cap-synthesizing complex associated with RNA polymerase. This enzymatic complex catalyzes the chemical reactions that are required for mRNA capping. Synthesis proceeds as a multi-step biochemical reaction.

Splicing

Splicing is the process by which pre-mRNA is modified to remove certain stretches of non-coding sequences called introns; the stretches that remain include protein-coding sequences and are called exons. Sometimes pre-mRNA messages may be spliced in several different ways, allowing a single gene to encode multiple proteins. This process is called alternative splicing. Splicing is usually performed by an RNA-protein complex called the spliceosome, but some RNA molecules are also capable of catalyzing their own splicing.

Editing

In some instances, an mRNA will be edited, changing the nucleotide composition of that mRNA. An example in humans is the apolipoprotein B mRNA, which is edited in some tissues, but not others. The editing creates an early stop codon, which upon translation, produces a shorter protein.

Polyadenylation

Polyadenylation is the covalent linkage of a polyadenylyl moiety to a messenger RNA molecule. In eukaryotic organisms, most messenger RNA (mRNA) molecules are polyadenylated at the 3' end. The poly(A) tail and the protein bound to it aid in protecting mRNA from degradation by exonucleases. Polyadenylation is also important for transcription termination, export of the mRNA from the nucleus, and translation. mRNA can also be polyadenylated in prokaryotic organisms, where poly(A) tails act to facilitate, rather than impede, exonucleolytic degradation.

Polyadenylation occurs during and immediately after transcription of DNA into RNA. After transcription has been terminated, the mRNA chain is cleaved through the action of an endonuclease complex associated with RNA polymerase. After the mRNA has been cleaved, around 250 adenosine residues are added to the free 3' end at the cleavage site. This reaction is catalyzed by polyadenylate polymerase. Just as in alternative splicing, there can be more than one polyadenylation variant of a mRNA.

Transport

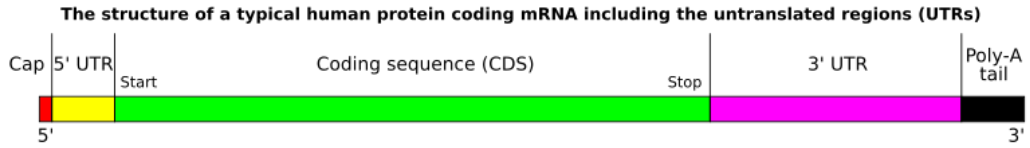
Another difference between eukaryotes and prokaryotes is mRNA transport. Because eukaryotic transcription and translation is compartmentally separated, eukaryotic mRNAs must be exported from the nucleus to the cytoplasm. Mature mRNAs are recognized by their processed modifications and then exported through the nuclear pore. In neurons mRNA must be transported from the soma to the dendrites where local translation occurs in response to external stimuli. Many messages are marked with so-called "zip codes" which targets their transport to a specific location.

Translation

Because prokaryotic mRNA does not need to be processed or transported, translation by the ribosome can begin immediately after the end of transcription. Therefore, it can be said that prokaryotic translation is *coupled* to transcription and occurs *co-transcriptionally*.

Eukaryotic mRNA that has been processed and transported to the cytoplasm (i.e. mature mRNA) can then be translated by the ribosome. Translation may occur at ribosomes free-floating in the cytoplasm, or directed to the endoplasmic reticulum by the signal recognition particle. Therefore, unlike in prokaryotes, eukaryotic translation *is not* directly coupled to transcription.

Structure



The structure of a mature eukaryotic mRNA. A fully processed mRNA includes a 5' cap, 5' UTR, coding region, 3' UTR, and poly(A) tail.

5' cap

The *5′ cap* is a modified guanine nucleotide added to the "front" (5' end) of the pre-mRNA using a 5'-5'-triphosphate linkage. This modification is critical for recognition and proper attachment of mRNA to the ribosome, as well as protection from 5' exonucleases. It may also be important for other essential processes, such as splicing and transport.

Coding regions

Coding regions are composed of codons, which are decoded and translated (in eukaryotes usually into one and in prokaryotes usually into several) proteins by the ribosome. Coding regions begin with the start codon and end with a stop codon. Generally, the start codon is an AUG triplet and the stop codon is UAA, UAG, or UGA. The coding regions tend to be stabilised by internal base pairs, this impedes degradation. In addition to being protein-coding, portions of coding regions may serve as regulatory sequences in the pre-mRNA as exonic splicing enhancers or exonic splicing silencers.

Untranslated regions

Untranslated regions (UTRs) are sections of the mRNA before the start codon and after the stop codon that are not translated, termed the five prime untranslated region (5' UTR) and three prime untranslated region (3' UTR), respectively. These regions are transcribed

with the coding region and thus are exonic as they are present in the mature mRNA. Several roles in gene expression have been attributed to the untranslated regions, including mRNA stability, mRNA localization, and translational efficiency. The ability of a UTR to perform these functions depends on the sequence of the UTR and can differ between mRNAs.

The stability of mRNAs may be controlled by the 5' UTR and/or 3' UTR due to varying affinity for RNA degrading enzymes called ribonucleases and for ancillary proteins that can promote or inhibit RNA degradation.

Translational efficiency, including sometimes the complete inhibition of translation, can be controlled by UTRs. Proteins that bind to either the 3' or 5' UTR may affect translation by influencing the ribosome's ability to bind to the mRNA. MicroRNAs bound to the 3' UTR also may affect translational efficiency or mRNA stability.

Cytoplasmic localization of mRNA is thought to be a function of the 3' UTR. Proteins that are needed in a particular region of the cell can actually be translated there; in such a case, the 3' UTR may contain sequences that allow the transcript to be localized to this region for translation.

Some of the elements contained in untranslated regions form a characteristic secondary structure when transcribed into RNA. These structural mRNA elements are involved in regulating the mRNA. Some, such as the SECIS element, are targets for proteins to bind. One class of mRNA element, the riboswitches, directly bind small molecules, changing their fold to modify levels of transcription or translation. In these cases, the mRNA regulates itself.

Poly(A) tail

The 3' poly(A) tail is a long sequence of adenine nucleotides (often several hundred) added to the 3' end of the pre-mRNA. This tail promotes export from the nucleus and translation, and protects the mRNA from degradation.

Monocistronic versus polycistronic mRNA

An mRNA molecule is said to be monocistronic when it contains the genetic information to translate only a single protein. This is the case for most of the eukaryotic mRNAs. On the other hand, polycistronic mRNA carries the information of several genes, which are translated into several proteins. These proteins usually have a related function and are grouped and regulated together in an operon. Most of the mRNA found in bacteria and archea are polycistronic. Dicistronic or bicistronic is the term used to describe an mRNA that encodes only two proteins.

mRNA circularization

In eukaryotes it is thought that mRNA molecules form circular structures due to an interaction between the cap binding complex and poly(A)-binding protein. Circularization is thought to promote recycling of ribosomes on the same message leading to efficient translation.

Degradation

Different mRNAs within the same cell have distinct lifetimes (stabilities). In bacterial cells, individual mRNAs can survive from seconds to more than an hour; in mammalian cells, mRNA lifetimes range from several minutes to days. The greater the stability of an mRNA, the more protein may be produced from that mRNA. The limited lifetime of mRNA enables a cell to alter protein synthesis rapidly in response to its changing needs. There are many mechanisms that lead to the destruction of a mRNA, some of which are described below.

Prokaryotic mRNA degradation

In prokaryotes the lifetime of mRNA is generally much shorter than in eukaryotes. Prokaryotes degrade messages by using a combination of ribonucleases, including endonucleases, 3' exonucleases, and 5' exonucleases. In some instances, small RNA molecules (sRNA) tens to hundreds of nucleotides long can stimulate the degradation of specific mRNAs by base pairing with complementary sequences and facilitating ribonuclease cleavage. It was recently shown that bacteria also have a sort of 5' cap consisting of a triphosphate on the 5' end. Removal of two of the phosphates leaves a 5' monophosphate, causing the message to be destroyed by the endonuclease RNase E.

Eukaryotic mRNA turnover

Inside eukaryotic cells there is a balance between the processes of translation and mRNA decay. Messages that are being actively translated are bound by ribosomes, the eukaryotic initiation factors eIF-4E and eIF-4G, and poly(A)-binding protein. eIF-4E and eIF-4G block the decapping enzyme (DCP2), and poly(A)-binding protein blocks the exosome complex, protecting the ends of the message. The balance between translation and decay is reflected in the size and abundance of cytoplasmic structures known as P-bodies The poly(A) tail of the mRNA is shortened by specialized exonucleases that are targeted to specific messenger RNAs by a combination of cis-regulatory sequences on the RNA and trans-acting RNA-binding proteins. Poly(A) tail removal is thought to disrupt the circular structure of the message and destabilize the cap binding complex. The message is then subject to degradation by either the exosome complex or the decapping complex. In this way, translationally inactive messages can be destroyed quickly, while active messages remain intact. The mechanism by which translation stops and the message is handed-off to decay complexes is not understood in detail.

AU-rich element decay

The presence of AU-rich elements in some mammalian mRNAs tends to destabilize those transcripts through the action of cellular proteins that bind these sequences and stimulate poly(A) tail removal. Loss of the poly(A) tail is thought to promote mRNA degradation by facilitating attack by both the exosome complex and the decapping complex. Rapid mRNA degradation via AU-rich elements is a critical mechanism for preventing the overproduction of potent cytokines such as tumor necrosis factor (TNF) and granulocyte-macrophage colony stimulating factor (GM-CSF). AU-rich elements also regulate the biosynthesis of proto-oncogenic transcription factors like c-Jun and c-Fos.

Nonsense mediated decay

Eukaryotic messages are subject to surveillance by nonsense mediated decay (NMD), which checks for the presence of premature stop codons (nonsense codons) in the message. These can arise via incomplete splicing, V(D)J recombination in the adaptive immune system, mutations in DNA, transcription errors, leaky scanning by the ribosome causing a frame shift, and other causes. Detection of a premature stop codon triggers mRNA degradation by 5' decapping, 3' poly(A) tail removal, or endonucleolytic cleavage.

Small interfering RNA (siRNA)

In metazoans, small interfering RNAs (siRNAs) processed by Dicer are incorporated into a complex known as the RNA-induced silencing complex or RISC. This complex contains an endonuclease that cleaves perfectly complementary messages to which the siRNA binds. The resulting mRNA fragments are then destroyed by exonucleases. siRNA is commonly used in laboratories to block the function of genes in cell culture. It is thought to be part of the innate immune system as a defense against double-stranded RNA viruses.

MicroRNA (miRNA)

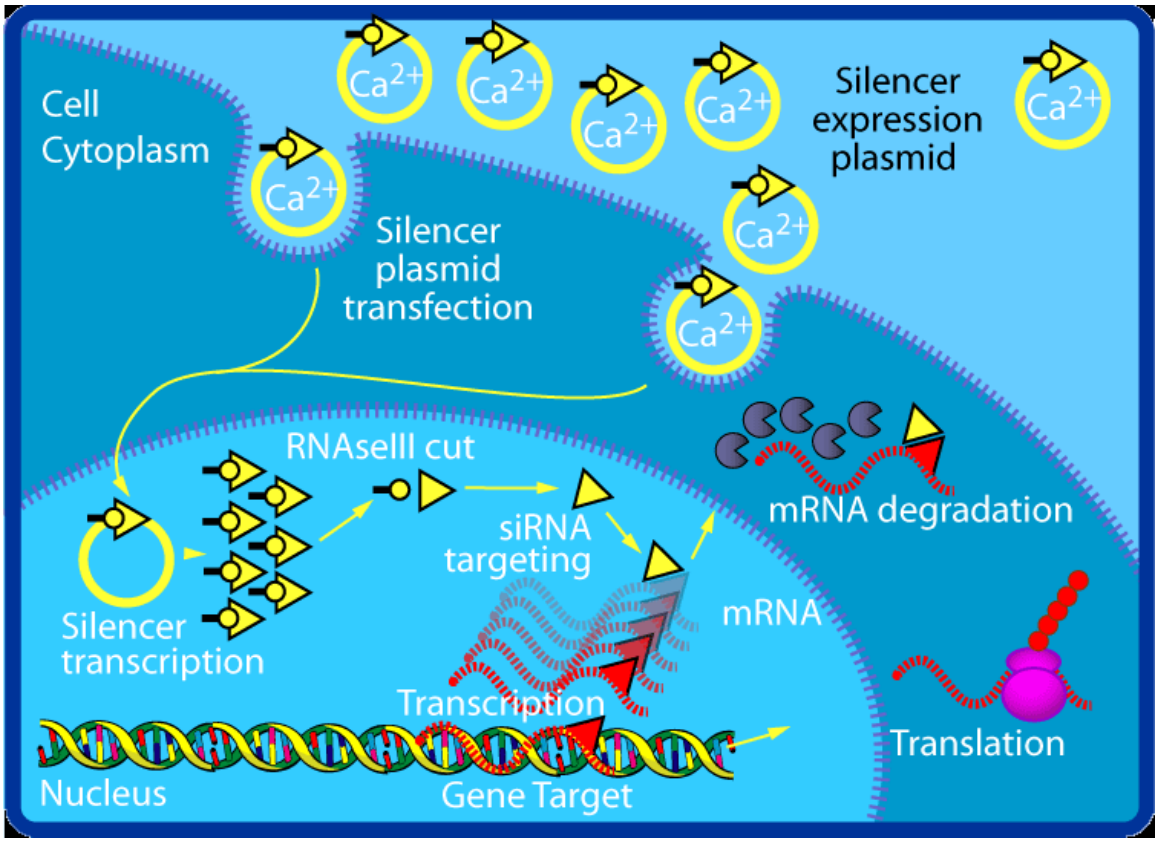
MicroRNAs (miRNAs) are small RNAs that typically are partially complementary to sequences in metazoan messenger RNAs. Binding of a miRNA to a message can repress translation of that message and accelerate poly(A) tail removal, thereby hastening mRNA degradation. The mechanism of action of miRNAs is the subject of active research.

Other decay mechanisms

There are other ways in which messages can be degraded, including non-stop decay, silencing by Piwi-interacting RNA (piRNA), and surely other means.

Chapter- 11

Small Interfering RNA



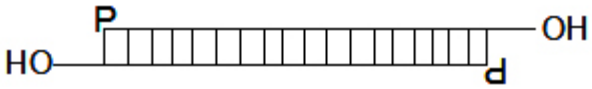
Mediating RNA interference in cultured mammalian cells

Small interfering RNA (**siRNA**), sometimes known as **short interfering RNA** or **silencing RNA**, is a class of double-stranded RNA molecules, 20-25 nucleotides in length, that play a variety of roles in biology. Most notably, siRNA is involved in the RNA interference (RNAi) pathway, where it interferes with the expression of a specific gene. In addition to their role in the RNAi pathway, siRNAs also act in RNAi-related pathways, e.g., as an antiviral mechanism or in shaping the chromatin structure of a genome; the complexity of these pathways is only now being elucidated.

siRNAs were first discovered by David Baulcombe's group at the Sainsbury Laboratory in Norwich, England, as part of post-transcriptional gene silencing (PTGS) in plants. The group published their findings in *Science* in a paper titled "A species of small antisense RNA in posttranscriptional gene silencing in plants". Shortly thereafter, in 2001, synthetic siRNAs were shown to be able to induce RNAi in mammalian cells by Thomas Tuschl, and colleagues in a paper published in *Nature*. This discovery led to a surge in interest in harnessing RNAi for biomedical research and drug development.

Structure

siRNAs have a well-defined structure: a short (usually 21-nt) double strand RNA (dsRNA) with 2-nt 3' overhangs on either end:



Schematic representation of a siRNA molecule: a ~19-21basepair RNA core duplex that is followed by a 2 nucleotide 3' overhang on each strand. OH: 3' hydroxyl; P: 5' phosphate.

Each strand has a 5' phosphate group and a 3' hydroxyl (-OH) group. This structure is the result of processing by dicer, an enzyme that converts either long dsRNAs or small hairpin RNAs into siRNAs. siRNAs can also be exogenously (artificially) introduced into cells by various transfection methods to bring about the specific knockdown of a gene of interest. Essentially any gene of which the sequence is known can thus be targeted based on sequence complementarity with an appropriately tailored siRNA. This has made siRNAs an important tool for gene function and drug target validation studies in the post-genomic era.

RNAi induction using siRNAs or their biosynthetic precursors

Dicer protein colored by protein domain.

Transfection of an exogenous siRNA can be problematic because the gene knockdown effect is only transient, particularly in rapidly dividing cells. One way of overcoming this challenge is to modify the siRNA in such a way as to allow it to be expressed by an appropriate vector, e.g., a plasmid. This is done by the introduction of a loop between the two strands, thus producing a single transcript, which can be processed into a functional siRNA. Such transcription cassettes typically use an RNA polymerase III promoter (e.g., U6 or H1), which usually directs the transcription of small nuclear RNAs (snRNAs) (U6 is involved in gene splicing; H1 is the RNase component of human RNase P). It is assumed (although not known for certain) that the resulting siRNA transcript is then processed by Dicer.

RNA activation

It has recently been found that dsRNA can also activate gene expression, a mechanism that has been termed "small RNA-induced gene activation" or RNAa. It has been shown that dsRNAs targeting gene promoters induce potent transcriptional activation of associated genes. RNAa was demonstrated in human cells using synthetic dsRNAs, termed "small activating RNAs" (saRNAs). It is currently not known whether RNAa is conserved in other organisms.

Challenges: avoiding nonspecific effects

Because RNAi intersects with a number of other pathways, it is not surprising that on occasion nonspecific effects are triggered by the experimental introduction of an siRNA. When a mammalian cell encounters a double-stranded RNA such as an siRNA, it may mistake it as a viral by-product and mount an immune response. Furthermore, because structurally related microRNAs modulate gene expression largely via incomplete complementarity base pair interactions with a target mRNA, the introduction of an siRNA may cause unintended off-targeting.

Innate immunity

Introduction of too much siRNA can result in nonspecific events due to activation of innate immune responses. Most evidence to date suggests that this is probably due to activation of the dsRNA sensor PKR, although retinoic acid inducible gene I (RIG-I) may also be involved. The induction of cytokines via toll-like receptor 7 (TLR7) has also been described. One promising method of reducing the nonspecific effects is to convert the siRNA into a microRNA. MicroRNAs occur naturally, and by harnessing this endogenous pathway it should be possible to achieve similar gene knockdown at comparatively low concentrations of resulting siRNAs. This should minimize nonspecific effects.

Off-targeting

Off-targeting is another challenge to the use of siRNAs as a gene knockdown tool. Here, genes with incomplete complementarity are inadvertently downregulated by the siRNA (effectively, the siRNA acts as a miRNA), leading to problems in data interpretation and potential toxicity. This, however, can be partly addressed by designing appropriate control experiments, and siRNA design algorithms are currently being developed to produce siRNAs free from off-targeting. Genome-wide expression analysis, e.g., by microarray technology, can then be used to verify this and further refine the algorithms. A 2006 paper from the laboratory of Dr. Khvorova implicates 6- or 7-basepair-long stretches from position 2 onward in the siRNA matching with 3'UTR regions in off-targeted genes.

Possible therapeutic applications and challenges

Given the ability to knock down essentially any gene of interest, RNAi via siRNAs has generated a great deal of interest in both basic and applied biology. There are an increasing number of large-scale RNAi screens that are designed to identify the important genes in various biological pathways. Because disease processes also depend on the activity of multiple genes, it is expected that in some situations turning off the activity of a gene with an siRNA could produce a therapeutic benefit.

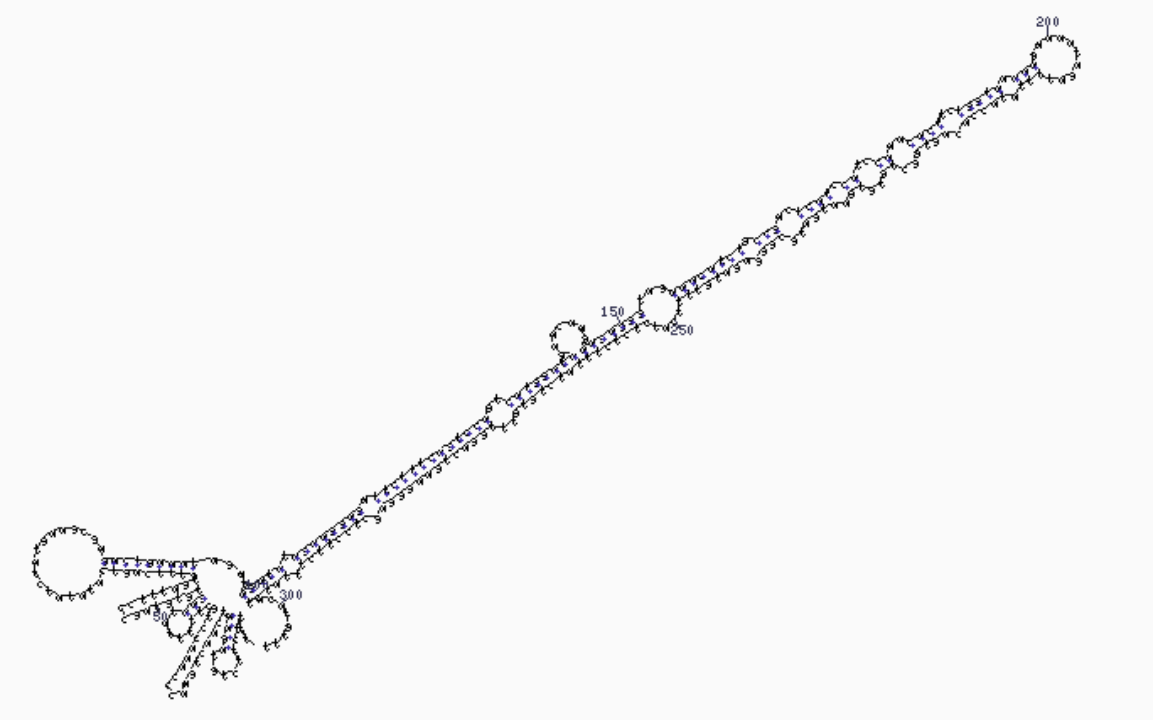
However, applying RNAi via siRNAs to living animals, especially humans, poses many challenges. Experimentally, siRNAs show different effectiveness in different cell types in a manner as yet poorly understood: some cells respond well to siRNAs and show a robust knockdown, whereas others show no such knockdown (even despite efficient transfection).

Phase I results of the first two therapeutic RNAi trials (indicated for age-related macular degeneration, aka AMD) reported at the end of 2005 that siRNAs are well tolerated and have suitable pharmacokinetic properties.

Proof of concept trials have indicated that Ebola-targeted siRNAs may be effective as post-exposure prophylaxis in humans, with 100% of non-human primates surviviing a lethal dose of Zaire Ebolavirus, the most lethal strain.

Chapter- 12

MicroRNA



The stem-loop secondary structure of a pre-microRNA from *Brassica oleracea*.

MicroRNAs (miRNAs) are short ribonucleic acid (RNA) molecules, on average only 22 nucleotides long and are found in all eukaryotic cells. miRNAs are post-transcriptional regulators that bind to complementary sequences on target messenger RNA transcripts (mRNAs), usually resulting in translational repression and gene silencing. The human genome may encode over 1000 miRNAs, which may target about 60% of mammalian genes and are abundant in many human cell types.

miRNAs show very different characteristics between plants and metazoans. In plants the miRNA complementarity to its mRNA target is nearly perfect, with no or few mismatched bases. In metazoans on the other hand miRNA complementarity is far from

perfect and one miRNA can target many different sites on the same mRNA or on many different mRNAs. Another difference is the location of target sites on mRNAs. In metazoans the miRNA target sites are in the three prime untranslated regions (3'UTR) of the mRNA. In plants targets can be located in the 3' UTR but are more often in the coding region itself. MiRNAs are well conserved in eukaryotic organism and are thought to be a vital and evolutionarily ancient component of genetic regulation.

The first miRNAs were characterized in the early 1990s, but miRNAs were not recognized as a distinct class of biologic regulators with conserved functions until the early 2000s. Since then, miRNA research has revealed multiple roles in negative regulation (transcript degradation and sequestering, translational suppression) and possible involvement in positive regulation (transcriptional and translational activation). By affecting gene regulation, miRNAs are likely to be involved in most biologic processes. Different sets of expressed miRNAs are found in different cell types and tissues.

Aberrant expression of miRNAs has been implicated in numerous disease states, and miRNA-based therapies are under investigation.

History

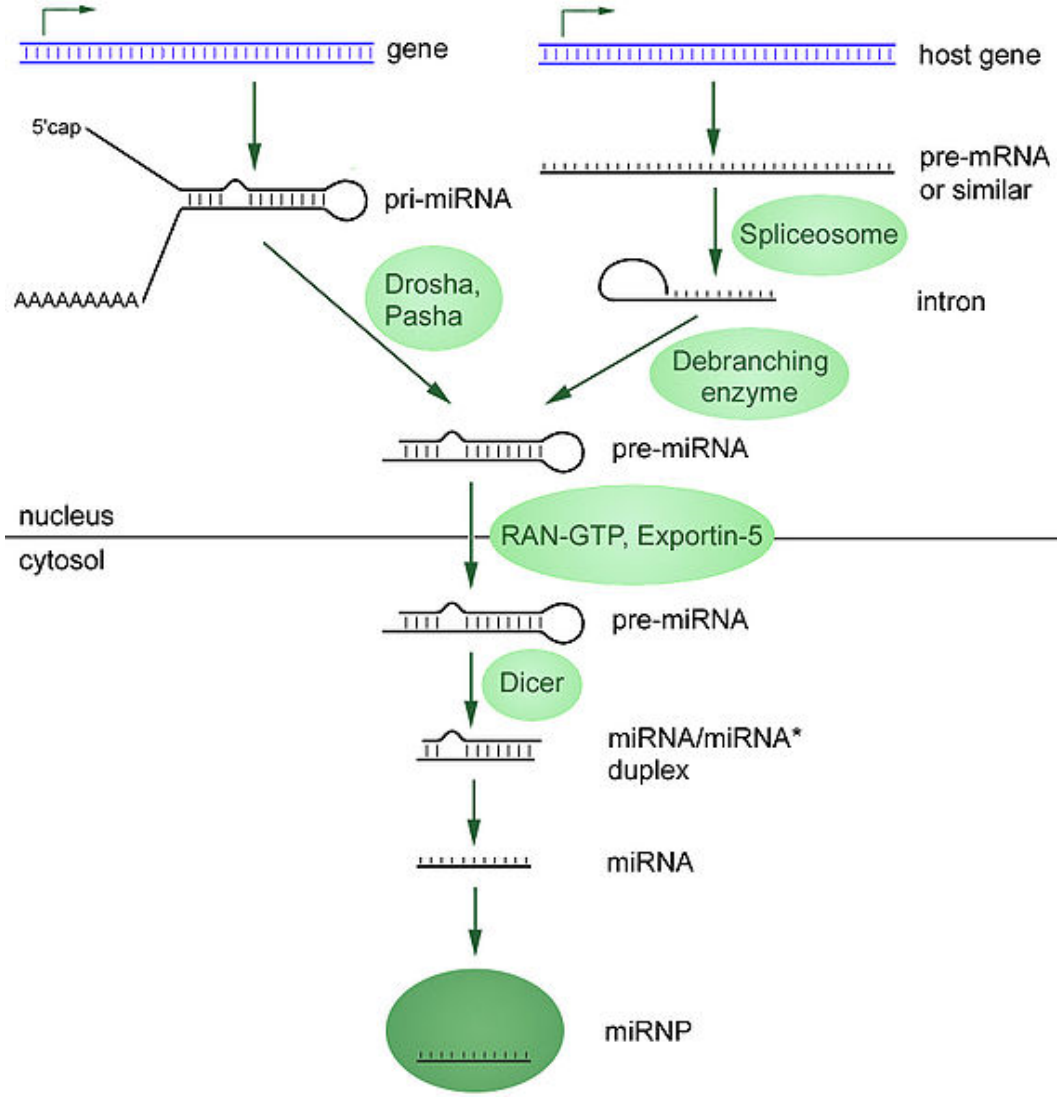
MicroRNAs were discovered in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum during a study of the gene *lin-14* in *C. elegans* development. They found that LIN-14 protein abundance was regulated by a short RNA product encoded by the *lin-4* gene. A 61 nucleotide precursor from *lin-4* gene matured to a 22 nucleotide RNA containing sequences partially complementary to multiple sequences in the 3’ UTR of the *lin-14* mRNA. This complementarity was sufficient and necessary to inhibit the translation of *lin-14* mRNA into LIN-14 protein. Retrospectively, the *lin-4* small RNA was the first microRNA to be identified, though at the time, it was thought to be a nematode idiosyncrasy. Only in 2000 was a second RNA characterized: let-7, which repressed *lin-41*, *lin-14*, *lin-28*, *lin-42*, and *daf-12* expression during developmental stage transitions in *C. elegans*. let-7 was soon found to be conserved in many species, indicating the existence of a wider phenomenon.

Nomenclature

Under a standard nomenclature system, names are assigned to experimentally confirmed miRNAs before publication of their discovery. The prefix "mir" is followed by a dash and a number, the latter often indicating order of naming. For example, mir-123 was named and likely discovered prior to mir-456. The uncapitalized "mir-" refers to the pre-miRNA, while a capitalized "miR-" refers to the mature form. miRNAs with nearly identical sequences bar one or two nucleotides are annotated with an additional lower case letter. For example, miR-123a would be closely related to miR-123b. Pre-miRNAs that lead to 100% identical mature miRNAs but that are located at different places in the genome are indicated with an additional dash-number suffix. For example, the pre-miRNAs hsa-mir-194-1 and hsa-mir-194-2 lead to an identical mature miRNA (hsa-miR-

194) but are located in different regions of the genome. Species of origin is designated with a three-letter prefix, e.g., hsa-miR-123 would be from human (*Homo sapiens*) and oar-miR-123 would be a sheep (*Ovis aries*) miRNA. Other common prefixes include 'v' for viral (miRNA encoded by a viral genome) and 'd' for *Drosophila* miRNA (a fruit fly commonly studied in genetic research). When two mature microRNAs originate from opposite arms of the same pre-miRNA, they are denoted with a -3p or -5p suffix. (In the past, this distinction was also made with 's' (sense) and 'as' (antisense)). When relative expression levels are known, an asterisk following the name indicates an miRNA expressed at low levels relative to the miRNA in the opposite arm of a hairpin. For example, miR-123 and miR-123* would share a pre-miRNA hairpin, but more miR-123 would be found in the cell.

Biogenesis



MicroRNAs are produced from either their own genes or from introns

Most microRNA genes are found in intergenic regions or in anti-sense orientation to genes and contain their own miRNA gene promoter and regulatory units. As much as 40% of miRNA genes may lie in the introns of protein and non-protein coding genes or even in exons. These are usually, though not exclusively, found in a sense orientation. and thus usually are regulated together with their host genes. Other miRNA genes showing a common promoter include the 42-48% of all miRNAs originating from polycistronic units containing 2-7 discrete loops from which mature miRNAs are processed, although this does not necessarily mean the mature miRNAs of a family will be homologous in structure and function. The promoters mentioned have been shown to have some similarities in their motifs to promoters of other genes transcribed by RNA polymerase II such as protein coding genes. The DNA template is not the final word on mature miRNA production: 6% of human miRNAs show RNA editing, the site-specific modification of RNA sequences to yield products different from those encoded by their DNA. This increases the diversity and scope of miRNA action beyond that implicated from the genome alone.

Transcription

miRNA genes are usually transcribed by RNA polymerase II (Pol II). The polymerase often binds to a promoter found near the DNA sequence encoding what will become the hairpin loop of the pre-miRNA. The resulting transcript is capped with a specially-modified nucleotide at the 5’ end, polyadenylated with multiple adenosines (a poly(A) tail), and spliced. The product, called a primary miRNA (pri-miRNA), may be hundreds or thousands of nucleotides in length and contain one or more miRNA stem loops. When a stem loop precursor is found in the 3’ UTR, a transcript may serve as a pri-miRNA and a mRNA. RNA polymerase III (Pol III) transcribes some miRNAs, especially those with upstream Alu sequences, transfer RNAs (tRNAs), and mammalian wide interspersed repeat (MWIR) promoter units.

Nuclear processing

A single pri-miRNA may contain from one to six miRNA precursors. These hairpin loop structures are composed of about 70 nucleotides each. Each hairpin is flanked by sequences necessary for efficient processing. The double-stranded RNA structure of the hairpins in a pri-miRNA is recognized by a nuclear protein known as DiGeorge Syndrome Critical Region 8 (DGCR8 or "Pasha" in invertebrates), named for its association with DiGeorge Syndrome. DGCR8 associates with the enzyme Drosha, a protein that cuts RNA, to form the "Microprocessor" complex. In this complex, DGCR8 orients the catalytic RNase III domain of Drosha to liberate hairpins from pri-miRNAs by cleaving RNA about eleven nucleotides from the hairpin base (two helical RNA turns into the stem). The resulting hairpin, known as a pre-miRNA (precursor-miRNA), has a two-nucleotide overhang at its 3’ end; it has 3' hydroxyl and 5' phosphate groups.

pre-miRNAs that are spliced directly out of introns, bypassing the Microprocessor complex, are known as "mirtrons." Originally thought to exist only in *Drosophila* and *C. elegans*, mirtrons have now been found in mammals.

Perhaps as many as 16% of pri-miRNAs may be altered through nuclear RNA editing. Most commonly, enzymes known as adenosine deaminases acting on RNA (ADARs) catalyze adenosine to inosine (A to I) transitions. RNA editing can halt nuclear processing (for example, of pri-miR-142, leading to degradation by the ribonuclease Tudor-SN) and alter downstream processes including cytoplasmic miRNA processing and target specificity (e.g., by changing the seed region of miR-376 in the central nervous system).

Nuclear export

pre-miRNA hairpins are exported from the nucleus in a process involving the nucleocytoplasmic shuttle Exportin-5. This protein, a member of the *karyopherin* family, recognizes a two-nucleotide overhang left by the RNase III enzyme Drosha at the 3' end of the pre-miRNA hairpin. Exportin-5-mediated transport to the cytoplasm is energy-dependent, using GTP bound to the Ran protein.

Cytoplasmic processing

In the cytoplasm, the pre-miRNA hairpin is cleaved by the RNase III enzyme Dicer. This endoribonuclease interacts with the 3' end of the hairpin and cuts away the loop joining the 3' and 5' arms, yielding an imperfect miRNA:miRNA* duplex about 22 nucleotides in length. Overall hairpin length and loop size influence the efficiency of Dicer processing, and the imperfect nature of the miRNA:miRNA* pairing also affects cleavage. Although either strand of the duplex may potentially act as a functional miRNA, only one strand is usually incorporated into the RNA-induced silencing complex (RISC) where the miRNA and its mRNA target interact.

Biogenesis in plants

miRNA biogenesis in plants differs from metazoan biogenesis mainly in the steps of nuclear processing and export. Instead of being cleaved by two different enzymes, once inside and once outside the nucleus, both cleavages of the plant miRNA is performed by a Dicer homolog, called Dicer-like1 (DL1). DL1 is only expressed in the nucleus of plant cells, which indicates that both reactions take place inside the nucleus. Before plant miRNA:miRNA* duplexes are transported out of the nucleus its 3' overhangs are methylated by a RNA methyltransferaseprotein called Hua-Enhancer1 (HEN1). The duplex is then transported out of the nucleus to the cytoplasm by a protein called Hasty (HST), an Exportin 5 homolog, where they disassemble and the mature miRNA is incorporated into the RISC.

The RNA-induced silencing complex

The mature miRNA is part of an active RNA-induced silencing complex (RISC) containing Dicer and many associated proteins. RISC is also known as a microRNA ribonucleoprotein complex (miRNP); RISC with incorporated miRNA is sometimes referred to as "miRISC."

Dicer processing of the pre-miRNA is thought to be coupled with unwinding of the duplex. Generally, only one strand is incorporated into the miRISC, selected on the basis of its thermodynamic instability and weaker base-pairing relative to the other strand. The position of the stem-loop may also influence strand choice. The other strand, called the passenger strand due to its lower levels in the steady state, is denoted with an asterisk (*) and is normally degraded. In some cases, both strands of the duplex are viable and become functional miRNA that target different mRNA populations.

Members of the argonaute (Ago) protein family are central to RISC function. Argonautes are needed for miRNA-induced silencing and contain two conserved RNA binding domains: a PAZ domain that can bind the single stranded 3’ end of the mature miRNA and a PIWI domain that structurally resembles ribonuclease-H and functions to interact with the 5’ end of the guide strand. They bind the mature miRNA and orient it for interaction with a target mRNA. Some argonautes, for example human Ago2, cleave target transcripts directly; argonautes may also recruit additional proteins to achieve translational repression. The human genome encodes eight argonaute proteins divided by sequence similarities into two families: AGO (with four members present in all mammalian cells and called E1F2C/hAgo in humans), and PIWI (found in the germ line and hematopoietic stem cells).

Additional RISC components include TRBP [human immunodeficiency virus (HIV) transactivating response RNA (TAR) binding protein], PACT (protein activator of the interferon induced protein kinase (PACT), the SMN complex, fragile X mental retardation protein (FMRP), and Tudor staphylococcal nuclease-domain-containing protein (Tudor-SN).

Mode of Silencing

Gene silencing may occur either via mRNA degradation or preventing mRNA from being translated. It has been demonstrated that if there is complete complementation between the miRNA and target mRNA sequence, Ago2 can cleave the mRNA and lead to direct mRNA degradation. Yet, if there isn't complete complementation the silencing is achieved by preventing translation.

miRNA turnover

Turnover of mature miRNA is needed for rapid changes in miRNA expression profiles. During miRNA maturation in the cytoplasm, uptake by the Argonaute protein is thought to stabilize the guide strand, while the opposite (* or "passenger") strand is preferentially destroyed. In what has been called a "Use it or lose it" strategy, Argonaute may preferentially retain miRNAs with many targets over miRNAs with few or no targets, leading to degradation of the non-targeting molecules.

Decay of mature miRNAs in animals is mediated by the 5′-to-3′ exoribonuclease XRN2, also known as Rat1p. In plants, SDN (small RNA degrading nuclease) family members

degrade miRNAs in the opposite (3'-to-5') direction. Similar enzymes are encoded in animal genomes, but their roles have not yet been described.

Several miRNA modifications affect miRNA stability. As indicated by work in the model organism *Arabidopsis thaliana* (thale cress), mature plant miRNAs appear to be stabilized by the addition of methyl moieties at the 3' end. The 2'-O-conjugated methyl groups block the addition of uracil (U) residues by uridyltransferase enzymes, a modification that may be associated with miRNA degradation. However, uridylation may also protect some miRNAs; the consequences of this modification are incompletely understood. Uridylation of some animal miRNAs has also been reported. Both plant and animal miRNAs may be altered by addition of adenine (A) residues to the 3' end of the miRNA. An extra A added to the end of mammalian miR-122, a liver-enriched miRNA important in Hepatitis C, stabilizes the molecule, and plant miRNAs ending with an adenine residue have slower decay rates.

Cellular functions

The function of miRNAs appears to be in gene regulation. For that purpose, a miRNA is complementary to a part of one or more messenger RNAs (mRNAs). Animal miRNAs are usually complementary to a site in the 3' UTR whereas plant miRNAs are usually complementary to coding regions of mRNAs. Perfect or near perfect base pairing with the target RNA promotes cleavage of the RNA. This is the primary mode of plant microRNAs. In animals, microRNAs more often only partially base pair and inhibit protein translation of the target mRNA (this exists in plants as well but is less common). MicroRNAs that are partially complementary to the target can also speed up deadenylation, causing mRNAs to be degraded sooner. For partially complementary microRNA to recognise their targets, the nucleotides 2–7 of the miRNA ('seed region') still have to be perfectly complementary. miRNAs occasionally also causes histone modification and DNA methylation of promoter sites and therefore affecting the expression of targeted genes.

Animal microRNAs target in particular developmental genes. In contrast, genes involved in functions common to all cells, such as gene expression, have very few microRNA target sites and seem to be under selection to avoid targeting by microRNAs.

dsRNA can also activate gene expression, a mechanism that has been termed "small RNA-induced gene activation" or RNAa. dsRNAs targeting gene promoters can induce potent transcriptional activation of associated genes. This was demonstrated in human cells using synthetic dsRNAs termed small activating RNAs (saRNAs), but has also been demonstrated for endogenous microRNA.

Evolution

MicroRNAs are significant phylogenetic markers because of their astonishingly low rate of evolution. Their origin may have permitted the development of morphological innovation, and by making gene expression more specific and 'fine-tunable', permitted the

genesis of complex organs and perhaps, ultimately, complex life. Indeed, rapid bursts of morphological innovation are generally associated with a high rate of microRNA accumulation.

MicroRNAs originate predominantly by the random formation of hairpins in "non-coding" sections of DNA (i.e. introns or intergene regions), but also by the duplication and modification of existing microRNAs. The rate of evolution (i.e. nucleotide substitution) in recently-originated microRNAs is comparable to that elsewhere in the non-coding DNA, implying evolution by neutral drift; however, older microRNAs have a much lower rate of change (often less than one substitution per hundred million years), suggesting that once a microRNA gains a function it undergoes extreme purifying selection. At this point, a microRNA is rarely lost from an animal's genome, although microRNAs which are more recently derived (and thus presumably non-functional) are frequently lost. This makes them a valuable phylogenetic marker, and they are being looked upon as a possible solution to such outstanding phylogenetic problems as the relationships of arthropods.

MicroRNAs feature in the genomes of most eukaryotic organisms, from the brown algae to the metazoa. Across all species, in excess of 5000 had been identified by March 2010. Whilst short RNA sequences (50 – hundreds of base pairs) of a broadly comparable function occur in bacteria, bacteria lack true microRNAs.

Experimental detection and manipulation of miRNA

MicroRNA expression can be quantified in a two-step polymerase chain reaction process of modified RT-PCR followed by quantitative real-time PCR. Variations of this method achieve absolute or relative quantification. miRNAs can also be hybridized to microarrays, slides or chips with probes to hundreds or thousands of miRNA targets, so that relative levels of miRNAs can be determined in different samples. MicroRNAs can be both discovered and profiled by high-throughput sequencing methods. The activity of an miRNA can be experimentally inhibited using a locked nucleic acid (LNA) oligo, a Morpholino oligo or a 2'-O-methyl RNA oligo. MicroRNA maturation can be inhibited at several points by steric-blocking oligos. The miRNA target site of an mRNA transcript can also be blocked by a steric-blocking oligo. Additionally, a specific miRNA can be silenced by a complementary antagomir. For the “in situ” detection of miRNA, the use of LNA is currently the only efficient method. The locked conformation of LNA results in enhanced hybridization properties and increases sensitivity and selectivity, making it ideal for detection of short miRNA.

miRNA and disease

Just as miRNA is involved in the normal functioning of eukaryotic cells, so has dysregulation of miRNA been associated with disease. A manually curated, publicly available database miR2Disease documents known relationships between miRNA dysregulation and human disease.

miRNA and cancer

Several miRNAs have been found to have links with some types of cancer.

A study of mice altered to produce excess c-Myc — a protein with mutated forms implicated in several cancers — shows that miRNA has an effect on the development of cancer. Mice that were engineered to produce a surplus of types of miRNA found in lymphoma cells developed the disease within 50 days and died two weeks later. In contrast, mice without the surplus miRNA lived over 100 days. Leukemia can be caused by the insertion of a viral genome next to the 17-92 array of microRNAs leading to increased expression of this microRNA.

Another study found that two types of miRNA inhibit the E2F1 protein, which regulates cell proliferation. miRNA appears to bind to messenger RNA before it can be translated to proteins that switch genes on and off.

By measuring activity among 217 genes encoding miRNA, patterns of gene activity that can distinguish types of cancers can be discerned. miRNA signatures may enable classification of cancer. This will allow doctors to determine the original tissue type which spawned a cancer and to be able to target a treatment course based on the original tissue type. miRNA profiling has already been able to determine whether patients with chronic lymphocytic leukemia had slow growing or aggressive forms of the cancer.

Transgenic mice that over-express or lack specific miRNAs have provided insight into the role of small RNAs in various malignancies.

A novel miRNA-profiling based screening assay for the detection of early-stage colorectal cancer has been developed and is currently in clinical trials. Early results showed that blood plasma samples collected from patients with early, resectable (Stage II) colorectal cancer could be distinguished from those of sex-and age-matched healthy volunteers. Sufficient selectivity and specificity could be achieved using small (less than 1 mL) samples of blood. The test has potential to be a cost-effective, non-invasive way to identify at-risk patients who should undergo colonoscopy.

miRNA and heart disease

The global role of miRNA function in the heart has been addressed by conditionally inhibiting miRNA maturation in the murine heart, and has revealed that miRNAs play an essential role during its development. miRNA expression profiling studies demonstrate that expression levels of specific miRNAs change in diseased human hearts, pointing to their involvement in cardiomyopathies. Furthermore, studies on specific miRNAs in animal models have identified distinct roles for miRNAs both during heart development and under pathological conditions, including the regulation of key factors important for cardiogenesis, the hypertrophic growth response, and cardiac conductance.

miRNA and the nervous system

miRNAs appear to regulate the nervous system. Neural miRNAs are involved at various stages of synaptic development, including dendritogenesis (involving miR-132, miR-134 and miR-124), synapse formation and synapse maturation (where miR-134 and miR-138 are thought to be involved). Some studies find altered miRNA expression in schizophrenia.

miRNA and non-coding RNAs

When the human genome project mapped its first chromosome in 1999, it was predicted the genome would contain over 100,000 protein coding genes. However, only around 20,000 were eventually identified (International Human Genome Sequencing Consortium, 2004). Since then, the advent of bioinformatics approaches combined with genome tiling studies examining the transcriptome, systematic sequencing of full length cDNA libraries, and experimental validation (including the creation of miRNA derived antisense oligonucleotides called antagomirs) have revealed that many transcripts are non protein-coding RNA, including several snoRNAs and miRNAs.