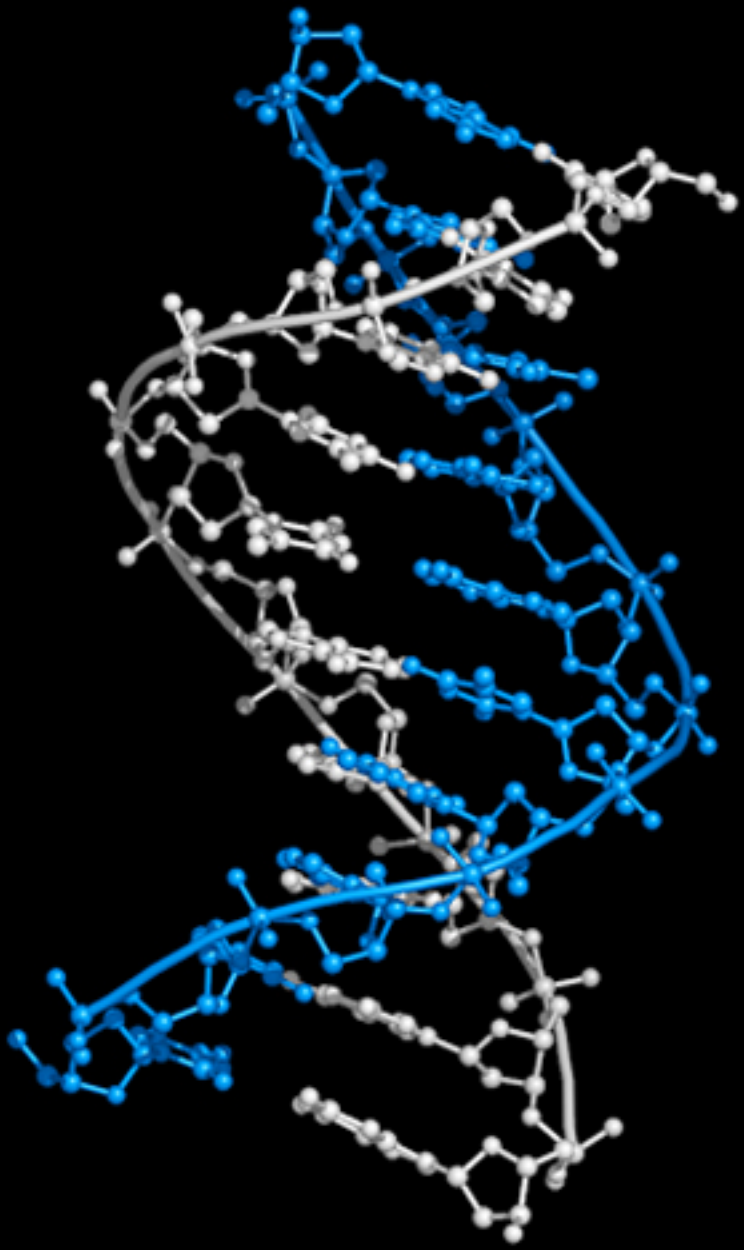
Nucleic Acid Structure



Skylar Ortega

First Edition, 2012

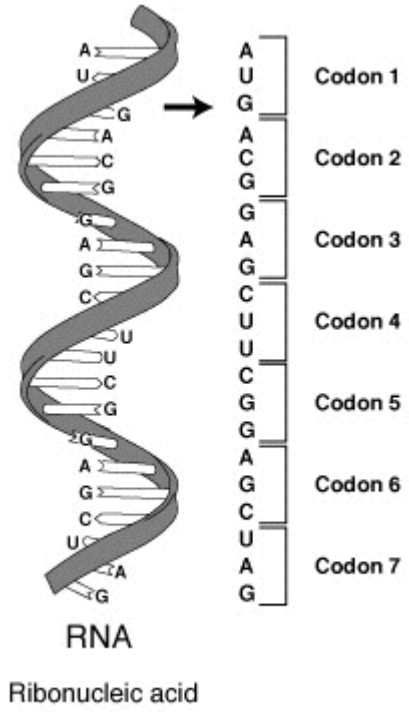
ISBN 978-81-323-3204-6

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 Sequence



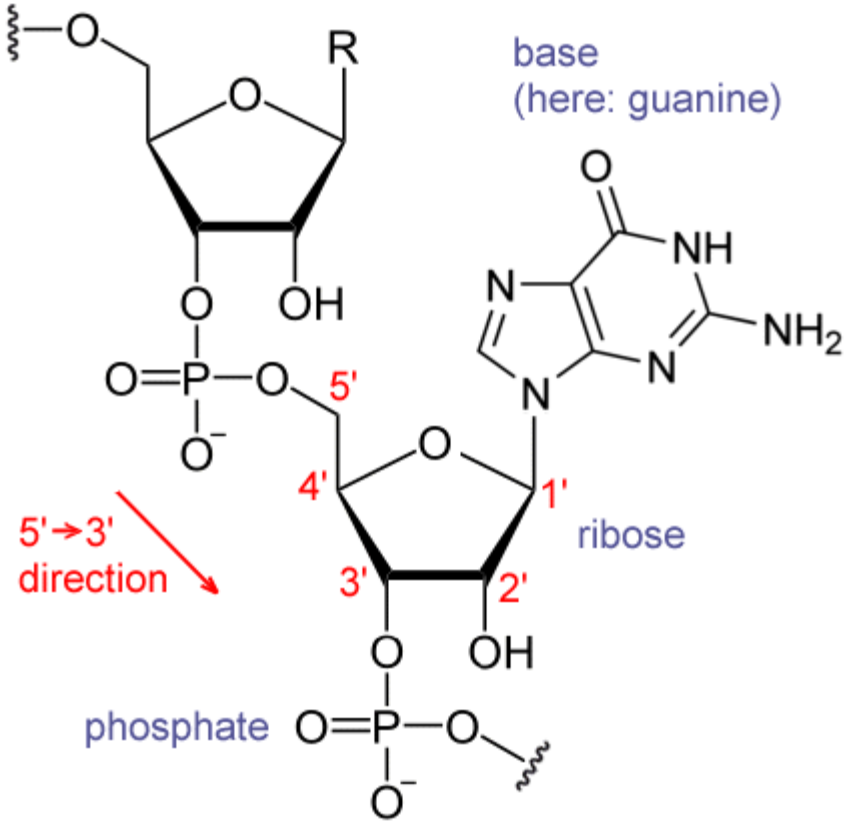
A series of codons in part of a mRNA molecule. Each codon consists of three nucleotides, usually representing a single amino acid.

The **sequence** or **primary structure of a nucleic acid** is the composition of atoms that make up the nucleic acid and the chemical bonds that those atoms. Because nucleic acids, such as DNA and RNA, are unbranched polymers, this specification is equivalent to specifying the sequence of nucleotides that comprise the molecule. This sequence is written as a succession of letters representing a real or hypothetical nucleic acid 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 represent information. Biological DNA represents the information which directs the functions of a living thing. In that context, the term **genetic sequence** is often used. Sequences can be read from the biological raw material through DNA sequencing methods.

Nucleic acids also have a secondary structure and tertiary structure. Primary structure is sometimes mistakenly referred to as *primary sequence*. Conversely, there is 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

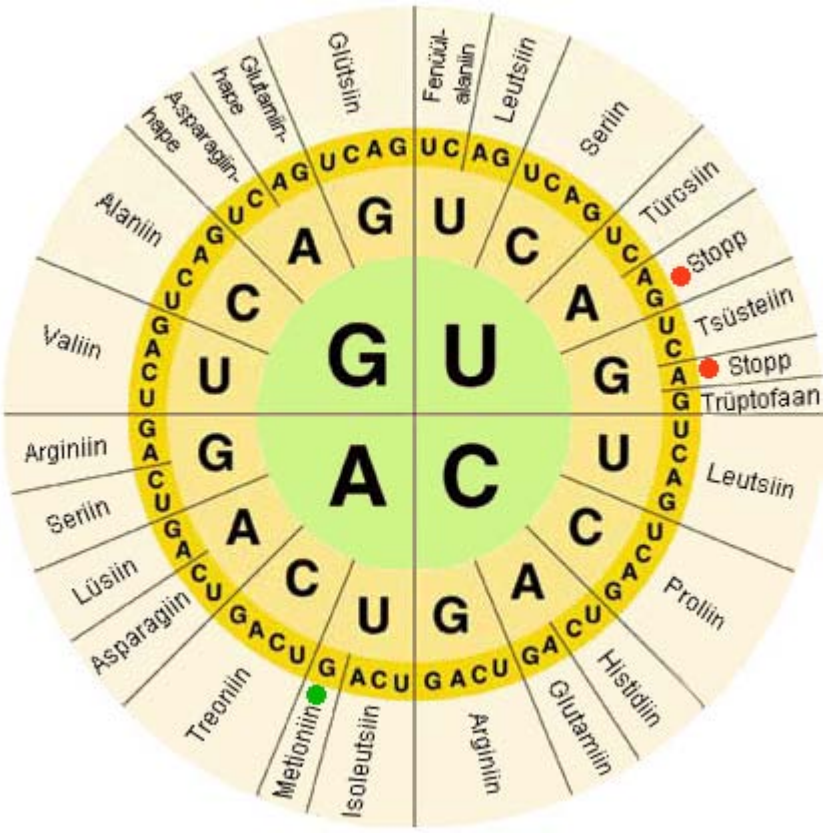
While A, T, C, and G represent a particular nucleotide at a position, there are also letters that represent ambiguity. Of all the molecules sampled, there is more than one kind of nucleotide at that position. The rules of the International Union of Pure and Applied Chemistry (IUPAC) are as follows:

- **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

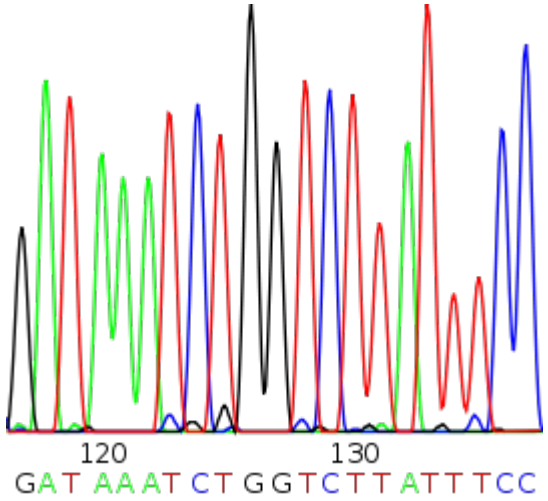


A depiction of the genetic code, by which the information contained in nucleic acids are translated into amino acid sequences in proteins.

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 the DNA of a living thing encodes the necessary information for that living thing to survive and reproduce. Therefore, determining the sequence is useful in fundamental research into why and how organisms live, as well as in applied subjects. Because of the importance of DNA to living things, knowledge of a DNA sequence may ne 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 therefore can only distinguish four bases. An inosine (created from adenosine during RNA editing) is read as a G, and 5-methyl-cytosine (created from cytosine by DNA methylation) is read as a C. With current technology, it is difficult to sequence small amounts of DNA, as the signal is too weak to measure. This is overcome by polymerase chain reaction (PCR) amplification.

Digital representation

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 variations of every gene, one inherited from their mother, the other 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. Usually, 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 due to 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 insertion or deletion mutations (indels) 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 2

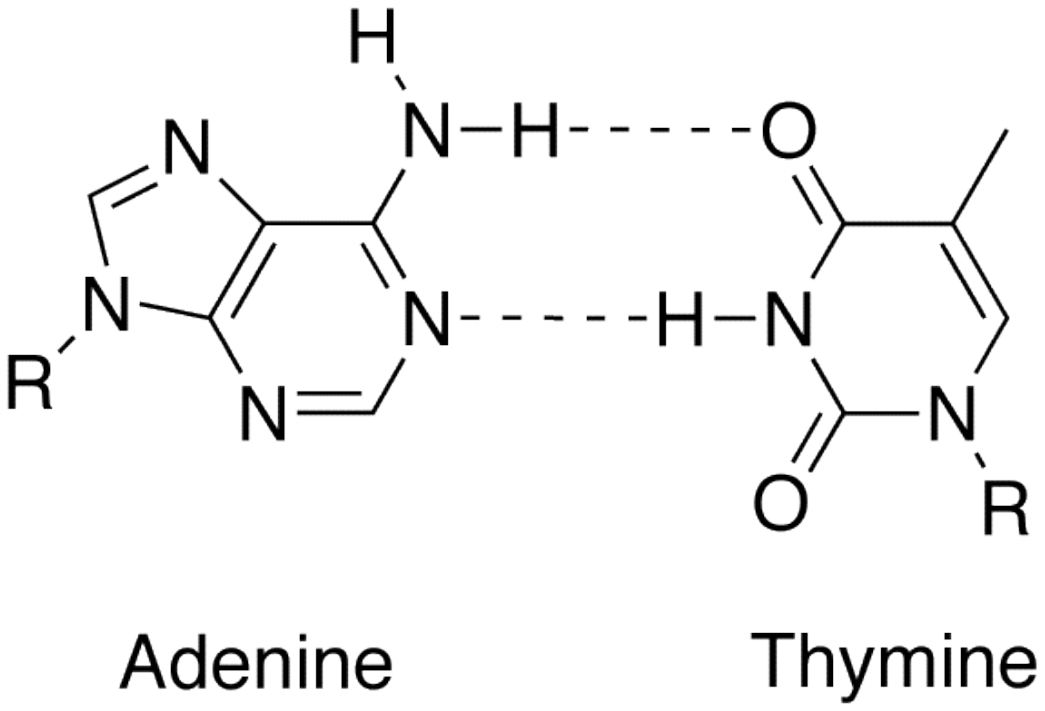
Nucleic Acid Secondary Structure

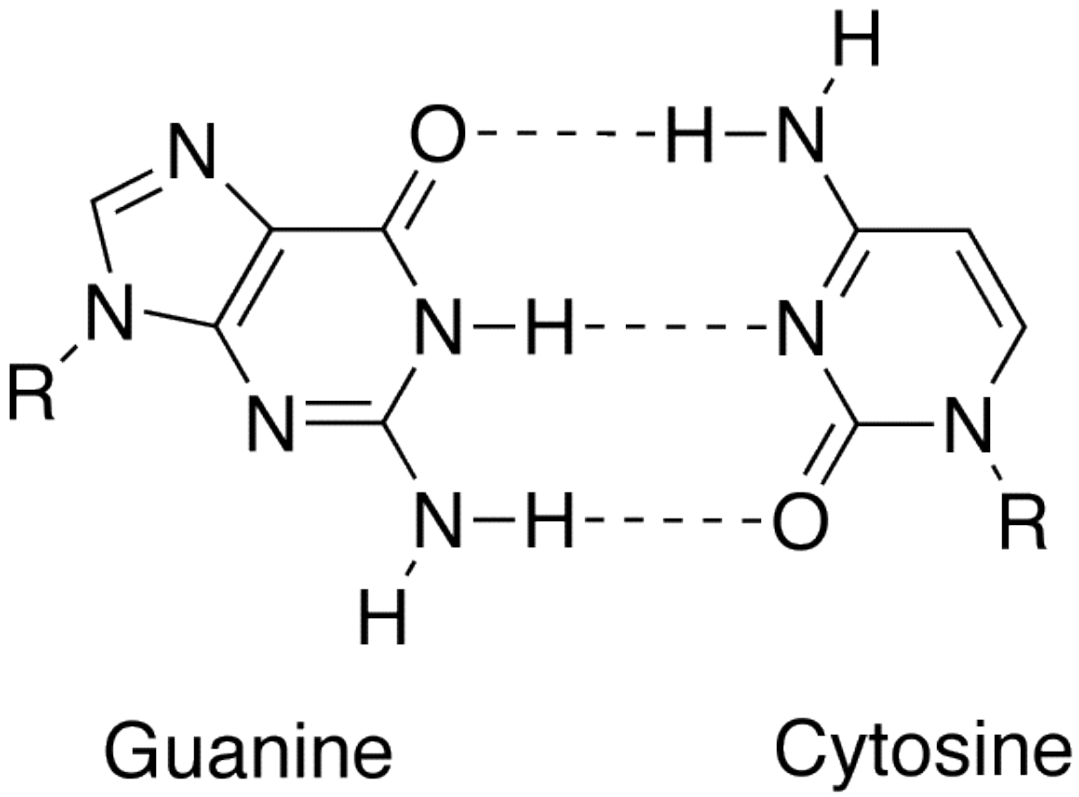
The **secondary structure of a nucleic acid molecule** refers to the basepairing interactions within a single molecule or set of interacting molecules, and can be represented as a list of bases which are paired in a nucleic acid molecule. The secondary structures of biological DNA's and RNA's tend to be different: biological DNA mostly exists as fully base paired double helices, while biological RNA is single stranded and often forms complicated base-pairing interactions due to its increased ability to form hydrogen bonds stemming from the extra hydroxyl group in the ribose sugar.

In a non-biological context, secondary structure is a vital consideration in the rational design of nucleic acid structures for DNA nanotechnology and DNA computing, since the pattern of basepairing ultimately determines the overall structure of the molecules.

Fundamental concepts

Base pairing





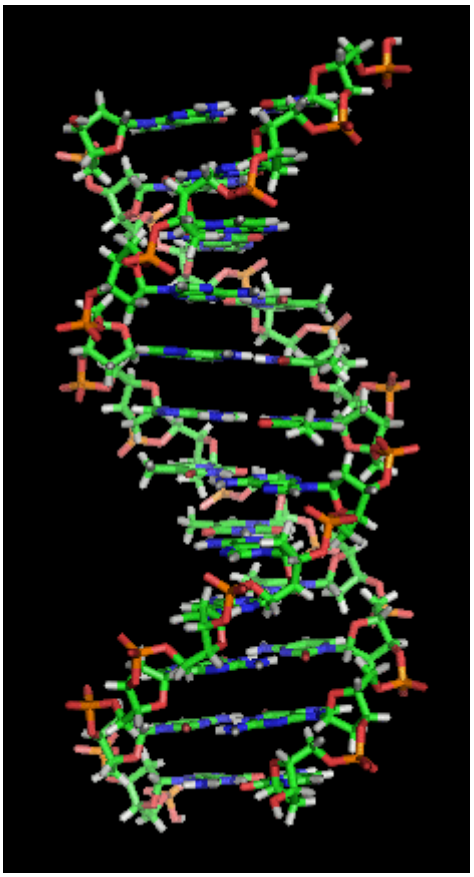
Top, an AT base pair demonstrating two intermolecular hydrogen bonds; *bottom*, a GC base pair demonstrating three intermolecular hydrogen bonds.

In molecular biology, two nucleotides on opposite complementary DNA or RNA strands that are connected via hydrogen bonds are called a base pair (often abbreviated bp). In the canonical Watson-Crick base pairing, adenine (A) forms a base pair with thymine (T), as and guanine (G) with cytosine (C) in DNA. In RNA, thymine is replaced by uracil (U). Alternate hydrogen bonding patterns, such as the wobble base pair and Hoogsteen base pair, also occur—particularly in RNA—giving rise to complex and functional tertiary structures. Importantly, pairing is the mechanism by which codons on messenger RNA molecules are recognized by anticodons on transfer RNA during protein translation. Some DNA- or RNA-binding enzymes can recognize specific base pairing patterns that identify particular regulatory regions of genes.

Hydrogen bonding is the chemical mechanism that underlies the base-pairing rules described above. Appropriate geometrical correspondence of hydrogen bond donors and acceptors allows only the "right" pairs to form stably. DNA with high GC-content is more stable than DNA with low GC-content, but contrary to popular belief, the hydrogen bonds do not stabilize the DNA significantly and stabilization is mainly due to stacking interactions.

The larger nucleobases, adenine and guanine, are members of a class of doubly-ringed chemical structures called purines; the smaller nucleobases, cytosine and thymine (and uracil), are members of a class of singly-ringed chemical structures called pyrimidines. Purines are only complementary with pyrimidines: pyrimidine-pyrimidine pairings are energetically unfavorable because the molecules are too far apart for hydrogen bonding to be established; purine-purine pairings are energetically unfavorable because the molecules are too close, leading to overlap repulsion. The only other possible pairings are GT and AC; these pairings are mismatches because the pattern of hydrogen donors and acceptors do not correspond. The GU wobble base pair, with two hydrogen bonds, does occur fairly often in RNA.

Nucleic acid hybridization



Two complementary regions of nucleic acid molcules will bind and from a double helical structure held together by base pairs.

Melting stability of base steps (B DNA)

Step	Melting ΔG /Kcal mol^{-1}
T A	-0.12
T G or **C A**	-0.78

C G	-1.44
A G or **C T**	-1.29
A A or **T T**	-1.04
A T	-1.27
G A or **T C**	-1.66
C C or **G G**	-1.97
A C or **G T**	-2.04
G C	-2.70

Hybridization is the process of complementary base pairs binding to form a double helix. Melting is the process by which the interactions between the strands of the double helix are broken, separating the two nucleic acid strands. These bonds are weak, easily separated by gentle heating, enzymes, or physical force. Melting occurs preferentially at certain points in the nucleic acid. **T** and **A** rich sequences are more easily melted than **C** and **G** rich regions. Particular base steps are also susceptible to DNA melting, particularly **T A** and **T G** base steps. These mechanical features are reflected by the use of sequences such as **TATAA** at the start of many genes to assist RNA polymerase in melting the DNA for transcription.

Strand separation by gentle heating, as used in PCR, is simple providing the molecules have fewer than about 10,000 base pairs (10 kilobase pairs, or 10 kbp). The intertwining of the DNA strands makes long segments difficult to separate. The cell avoids this problem by allowing its DNA-melting enzymes (helicases) to work concurrently with topoisomerases, which can chemically cleave the phosphate backbone of one of the strands so that it can swivel around the other. Helicases unwind the strands to facilitate the advance of sequence-reading enzymes such as DNA polymerase.

Secondary structure motifs

Nucleic acid secondary structure is generally divided into helices (contiguous base pairs), and various kinds of loops (unpaired nucleotides surrounded by helices). Frequently these elements, or combinations of them, can be further classified, for example, tetraloops, pseudoknots, and stem-loops.

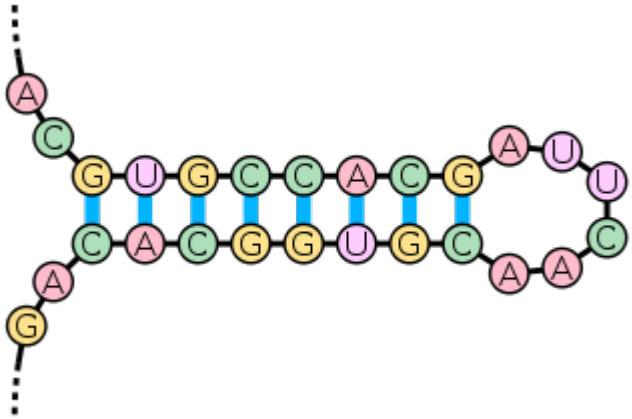
Double helix

The double helix is an important tertiary structure in nuleic acid molecules which is intimately connected with the molecule's secondary structure. A double helix is formed by regions of many consecutive base pairs.

The DNA double helix is a right-handed spiral polymer of nucleic acids, held together by nucleotides which base pair together. A single turn of the helix constitutes about ten nucleotides, and contains a major groove and minor groove, the major groove being wider than the minor groove. Given the difference in widths of the major groove and

minor groove, many proteins which bind to DNA do so through the wider major groove. Many double-helical forms are possible; for DNA the three biologically relevant forms are A-DNA, B-DNA, and Z-DNA, while RNA double helices have structures similar to the A form of DNA.

Stem-loop structures

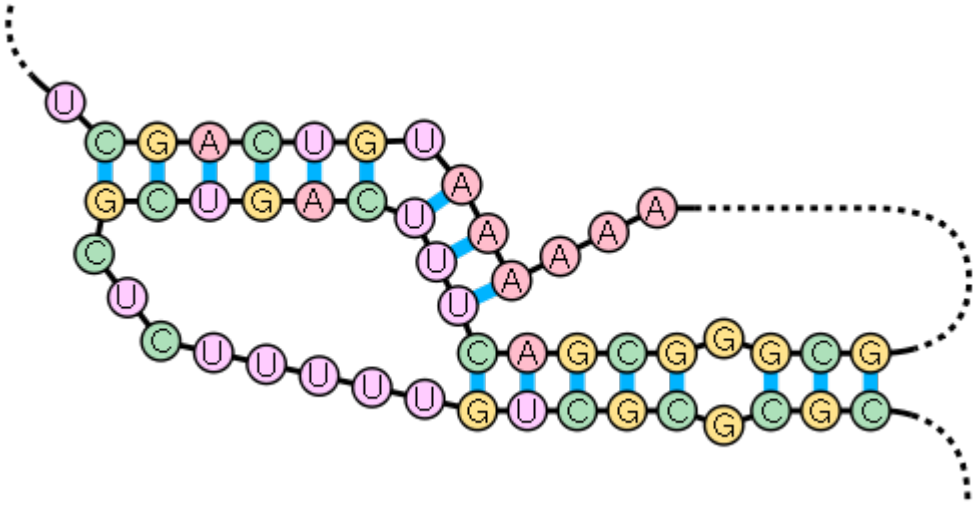


An example of an RNA stem-loop secondary structure

The secondary structure of nucleic acid molecules can often be uniquely decomposed into stems and loops. The stem-loop structure in which a base-paired helix ends in a short unpaired loop is extremely common and is a building block for larger structural motifs such as cloverleaf structures, which are four-helix junctions such as those found in transfer RNA. Internal loops (a short series of unpaired bases in a longer paired helix) and bulges (regions in which one strand of a helix has "extra" inserted bases with no counterparts in the opposite strand) are also frequent.

There are many secondary structure elements of functional importance to biological RNA's; some famous examples are the Rho-independent terminator stem-loops and the tRNA cloverleaf. There is a minor industry of researchers attempting to determine the secondary structure of RNA molecules. Approaches include both experimental and computational methods.

Pseudoknots



This example of a naturally occurring pseudoknot is found in the RNA component of human telomerase. Sequence from.

A pseudoknot is an RNA secondary structure containing at least two stem-loop structures in which half of one stem is intercalated between the two halves of another stem. Pseudoknots fold into knot-shaped three-dimensional conformations but are not true topological knots. The base pairing in pseudoknots is not well nested; that is, base pairs occur that "overlap" one another in sequence position. This makes the presence of general pseudoknots in RNA sequences impossible to predict by the standard method of dynamic programming, which uses a recursive scoring system to identify paired stems and consequently cannot detect non-nested base pairs with the most common algorithms. Limited subclasses of pseudoknots can be predicted using dynamic programs described in. Newer structure prediction techniques such as stochastic context-free grammars also do not take pseudoknots into account.

Several important biological processes rely on RNA molecules that form pseudoknots. For example, the RNA component of human telomerase contains a pseudoknot that is critical for activity. Though DNA can also form pseudoknots, they are generally not present in biological DNA.

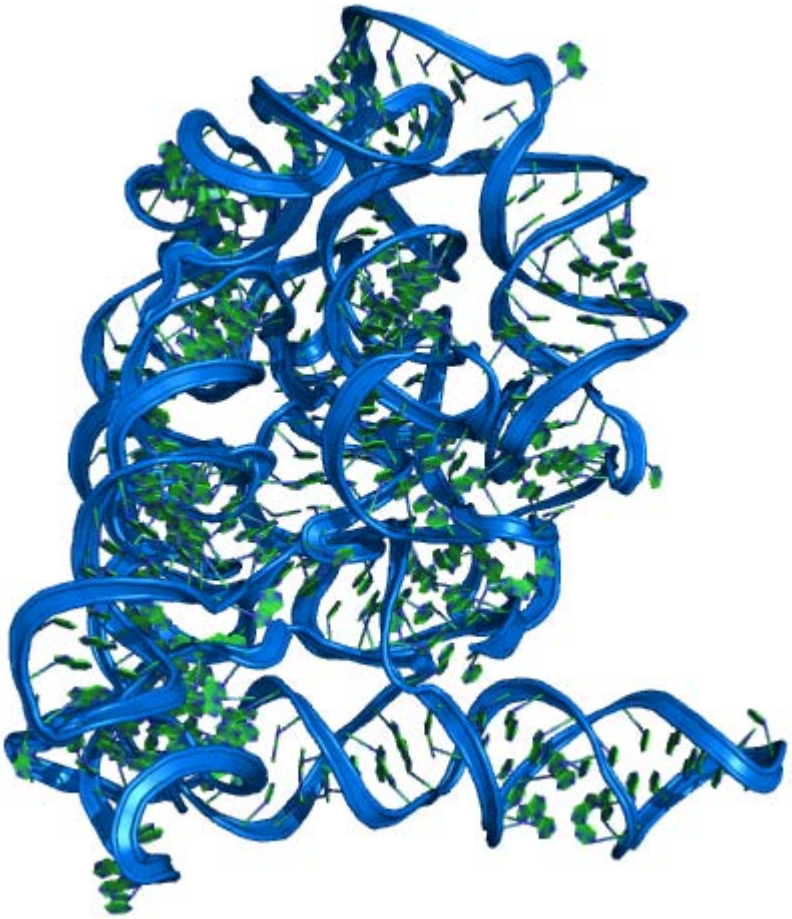
Secondary structure prediction

One application of bioinformatics uses predicted RNA secondary structures in searching a genome for noncoding but functional forms of RNA. For example, microRNAs have canonical long stem-loop structures interrupted by small internal loops. A general method of calculating probable RNA secondary structure is dynamic programming, although this has the disadvantage that it cannot detect pseudoknots or other cases in which base pairs are not fully nested. More general methods are based on stochastic context-free grammars. A web server that implements a type of dynamic programming is Mfold.

For many RNA molecules, the secondary structure is highly important to the correct function of the RNA — often more so than the actual sequence. This fact aids in the analysis of non-coding RNA sometimes termed "RNA genes". RNA secondary structure can be predicted with some accuracy by computer and many bioinformatics applications use some notion of secondary structure in analysis of RNA.

Chapter 3

Nucleic Acid Tertiary Structure

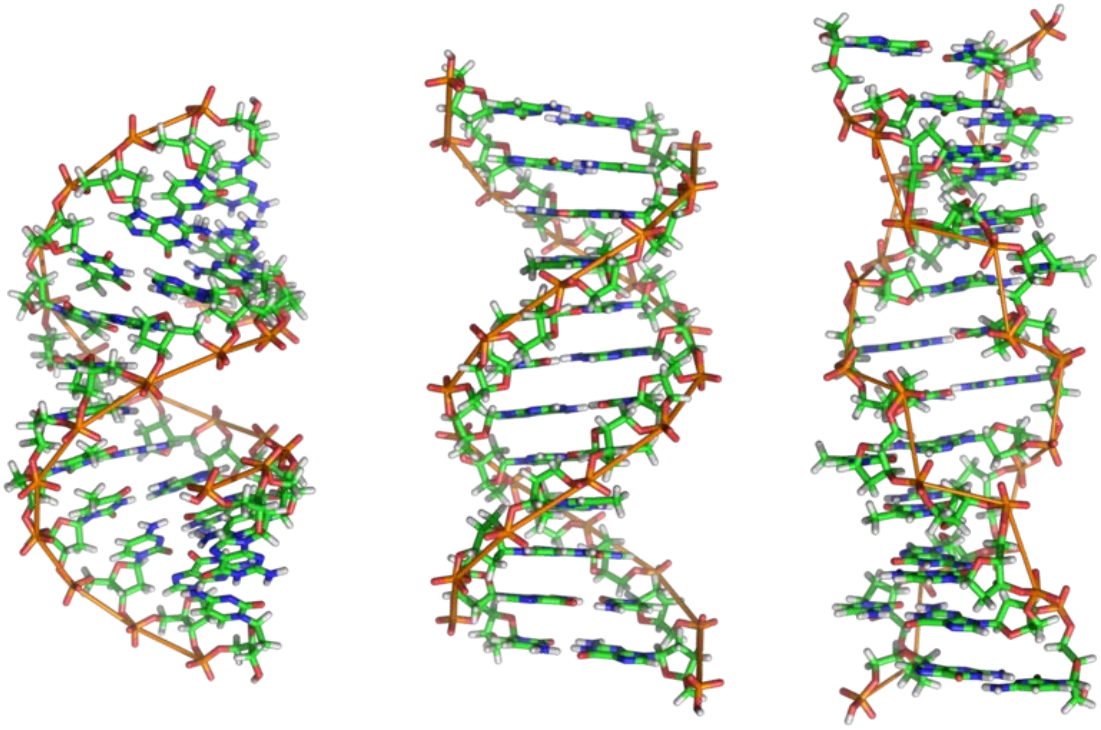


Example of a large catalytic RNA. The self-splicing group II intron from *Oceanobacillus iheyensis*.

The **tertiary structure of a nucleic acid** is its precise three-dimensional structure, as defined by the atomic coordinates. RNA and DNA molecules are capable of diverse

functions ranging from molecular recognition to catalysis. Such functions require a precise three-dimensional tertiary structure. While such structures are diverse and seemingly complex, they are composed of recurring, easily recognizable tertiary structure motifs that serve as molecular building blocks. Some of the most common motifs for RNA and DNA tertiary structure are described below, but it is important to remember that this information is based on a limited number of solved structures. Many more tertiary structural motifs will be revealed as new RNA and DNA molecules are structurally characterized.

Helical structures



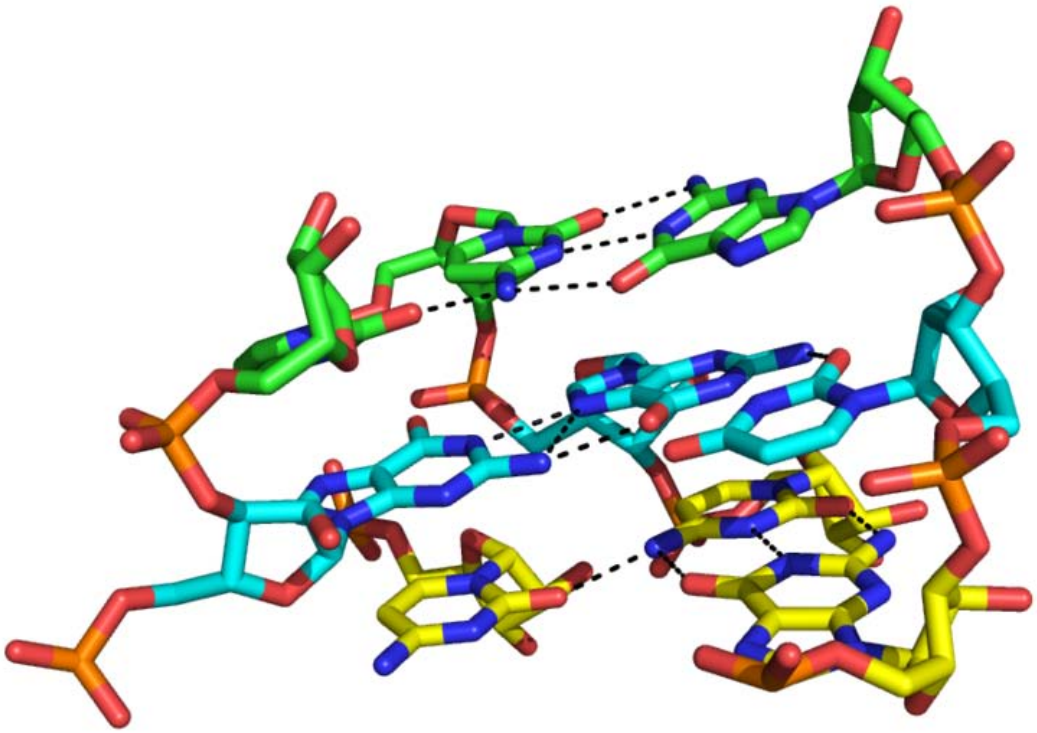
The structures of the A-, B-, and Z-DNA double helix structures.

Double helix

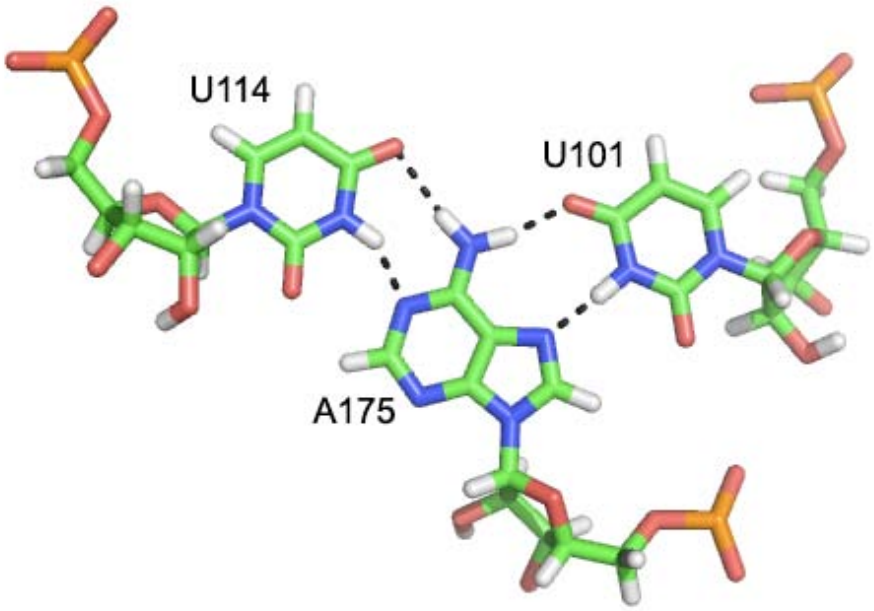
The double helix is the dominant tertiary structure for biological DNA, and is also a possible structure for RNA. Three DNA conformations are believed to be found in nature, A-DNA, B-DNA, and Z-DNA. The "B" form described by James D. Watson and Francis Crick is believed to predominate in cells. James D. Watson and Francis Crick described this structure as a double helix with a radius of 10 Å and pitch of 34 Å, making one complete turn about its axis every 10 bp of sequence. The double helix makes one complete turn about its axis every 10.4-10.5 base pairs in solution. This frequency of twist (known as the helical *pitch*) depends largely on stacking forces that each base exerts on its neighbours in the chain. Double-helical RNA adopts a conformation similar to the A-form structure.

Other conformations are possible; in fact, only the letters F, Q, U, V, and Y are now available to describe any new DNA structure that may appear in the future. However, most of these forms have been created synthetically and have not been observed in naturally occurring biological systems.

RNA triplexes



Major groove triples in the group II intron in *Oceanobacillus Iheyensis*. Each stacked layer is formed by one triplex with a different color scheme. Hydrogen bonds between triplexes are shown in black dashed lines. "N" atoms are colored in blue and "O" atoms in red. From top to bottom, the residues on the left side are G288, C289, and C377.



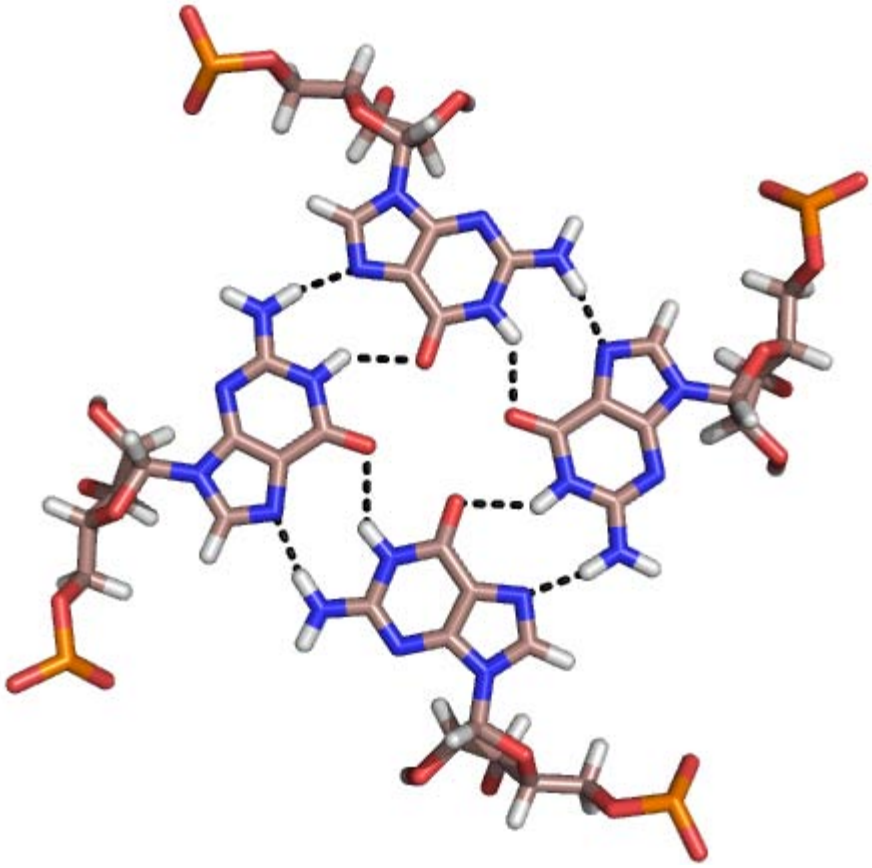
Close-up rendering of the U114:A175-U101 major groove (Hoogsteen base) triplex formed within the wild type pseudoknot of Human Telomerase RNA. Hydrogen bonds are shown in black dashed lines. "N" atoms are colored in blue and "o" atoms in red.

Major and minor groove triplexes

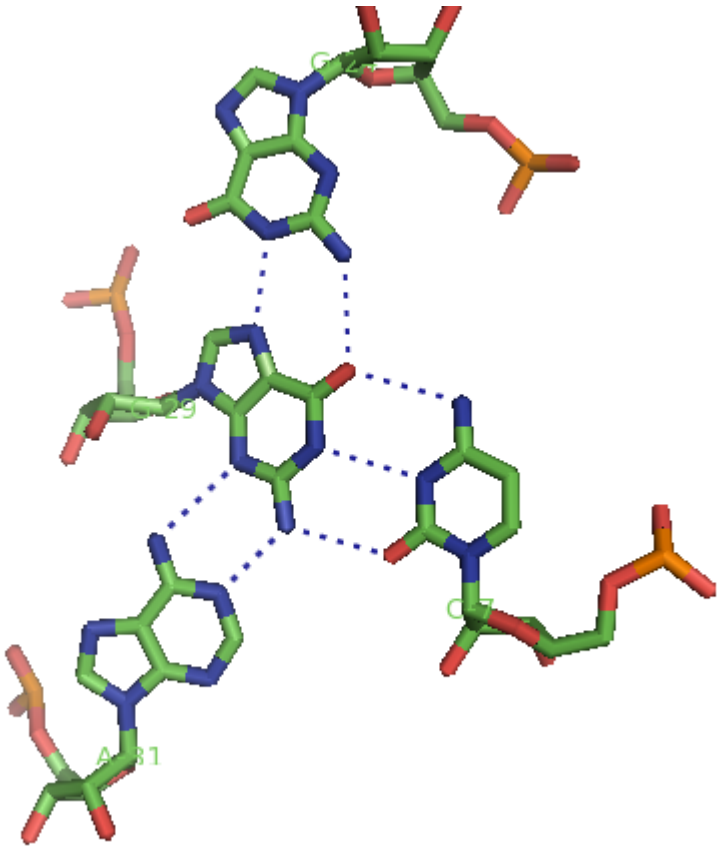
The minor groove triple is a ubiquitous RNA structural motif. Because interactions with the minor groove are often mediated by the 2'-OH of the ribose sugar, this RNA motif looks very different from its DNA equivalent. The most common example of a minor loop triple is the A-minor motif, or the insertion of adenosine bases into the minor groove. However, this motif is not restricted to adenosines, as other nucleobases have also been observed to interact with the RNA minor groove.

The minor groove presents a near-perfect complement for an inserted base. This allows for optimal van der Waals contacts, extensive hydrogen bonding and hydrophobic surface burial, and creates a highly energetically favorable interaction. Because minor groove triples are capable of stably packing a free loop and helix, they are key elements in the structure of large ribonucleotides, including the group I intron, the group II intron, and the ribosome.

Quadruplexes



Above:Typical Ring Structure of a Hoogsteen paired G-quartet.



Above: **Quadruplex** seen in crystal structure of Malachite Green RNA aptamer. G29 involved in major groove, minor groove, and Watson-Crick hydrogen-bonding with three other bases.

Although the major groove of standard A-form RNA is fairly narrow and therefore less available for triplex interaction than the minor groove, major groove triplex interactions can be observed in several RNA structures. These structures consist several combinations of base pair and Hoogsteen interactions. For example, the GGC triplex (GGC amino(N-2)-N-7, imino-carbonyl, carbonyl-amino(N-4); Watson-Crick) observed in the 50S ribosome, composed of a Watson-Crick type G-C pair and an incoming G which forms a pseudo-Hoogsteen network of hydrogen bonding interactions between both bases involved in the canonical pairing. Other notable examples of major groove triplexes include (i) the catalytic core of the group II intron shown in the figure at left (ii) a catalytically essential triple helix observed in human telomerase RNA and (iii) the SAM-II riboswitch.

Triple-stranded DNA is also possible from Hoogsteen or reversed Hoogsteen hydrogen bonds in the major groove of B-form DNA.

Quadruplexes

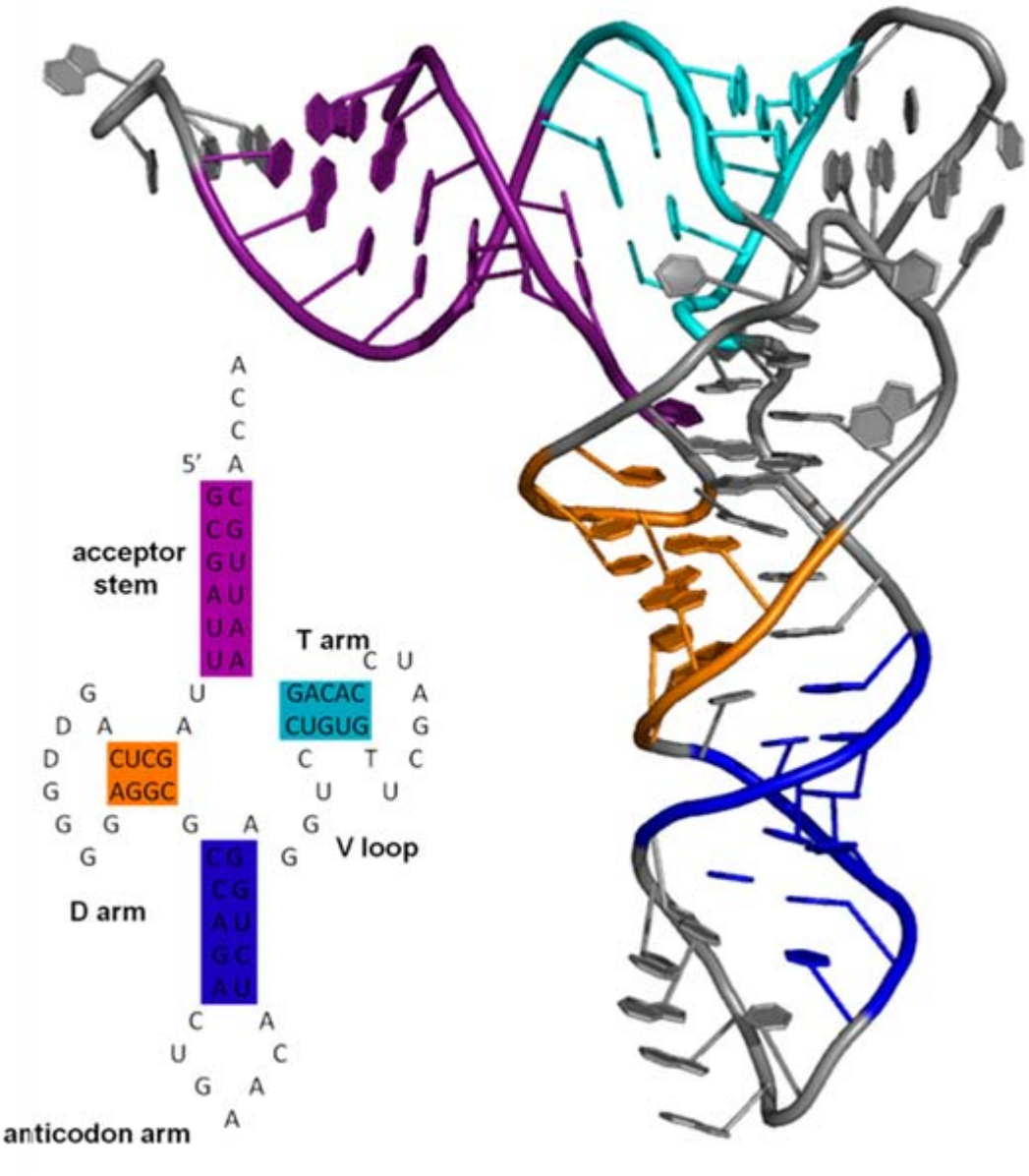
Besides double helices and the above-mentioned triplexes, RNA and DNA can both also form quadruple helices. There are diverse structures of RNA base quadruplexes. Four consecutive guanine residues can form a quadruplex in RNA by Hoogsteen hydrogen bonds to form a “Hoogsteen ring”. G-C and A-U pairs can also form base quadruplex with a combination of Watson-Crick pairing and noncanonical pairing in the minor groove.

The core of malachite green aptamer is also a kind of base quadruplex with a different hydrogen bonding pattern. The quadruplex can repeat several times consecutively, producing an immensely stable structure.

The unique structure of quadruplex regions in RNA may serve different functions in a biological system. Two important functions are the binding potential with ligands or proteins, and its ability to stabilize the whole tertiary structure of DNA or RNA. The strong structure can inhibit or modulate transcription and replication, such as in the telomeres of chromosomes and the UTR of mRNA. The base identity is important towards ligand binding. The G-quartet typically binds monovalent cations such as potassium, while other bases can bind numerous other ligands such as hypoxanthine in a U-U-C-U quadruplex.

Along with these functions, the G-quadruplex in the mRNA around the ribosome binding regions could serve as a regulator of gene expression in bacteria. There may be more interesting structures and functions yet to be discovered *in vivo*.

Coaxial stacking



Secondary (inset) and tertiary structure of tRNA demonstrating coaxial stacking.

Coaxial stacking, otherwise known as helical stacking, is a major determinant of higher order RNA tertiary structure. Coaxial stacking occurs when two RNA duplexes form a contiguous helix, which is stabilized by base stacking at the interface of the two helices. Coaxial stacking was noted in the crystal structure of tRNAPhe. More recently, coaxial stacking has been observed in higher order structures of many ribozymes, including many forms of the self-splicing group I and group II introns. Common coaxial stacking motifs include the kissing loop interaction and the pseudoknot. The stability of these interactions can be predicted by an adaptation of "Turner's rules".

In 1994, Walter and Turner determined the free energy contributions of nearest neighbor stacking interactions within a helix-helix interface by using a model system that created a helix-helix interface between a short oligomer and a four-nucleotide overhang at the end of a hairpin stem . Their experiments confirmed that the thermodynamic contribution of base-stacking between two helical secondary structures closely mimics the thermodynamics of standard duplex formation (nearest neighbor interactions predict the thermodynamic stability of the resulting helix). The relative stability of nearest neighbor interactions can be used to predict favorable coaxial stacking based on known secondary structure. Walter and Turner found that, on average, prediction of RNA structure improved from 67% to 74% accuracy when coaxial stacking contributions were included. Theories of coaxial stacking can be tested using the technique of helical fusion. This approach was used by Murphy and Cech to confirm a coaxial stacking interaction between the P4 and P6 helices within the catalytic center of the Tetrahymena group I intron.

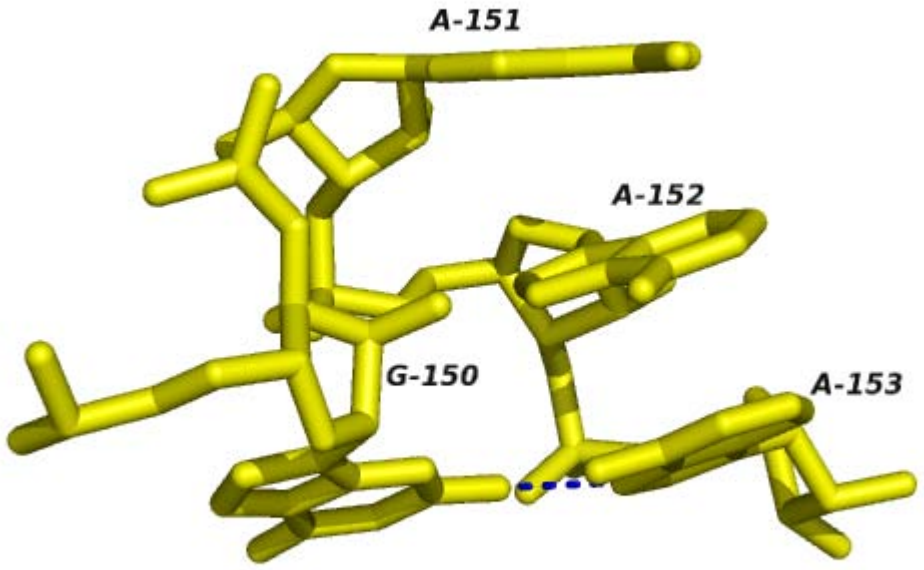
Most well-studied RNA tertiary structures contain examples of coaxial stacking. Some prominent examples are tRNA-Phe, group I introns, group II introns, and ribosomal RNAs. Crystal structures of tRNA revealed the presence of two extended helices that result from coaxial stacking of the amino-acid acceptor stem with the T-arm, and stacking of the D- and anticodon-arms. These interactions within tRNA orient the anticodon stem perpendicularly to the amino-acid stem, leading to the functional L-shaped tertiary structure. In group I introns, the P4 and P6 helices were shown to coaxially stack using a combination of biochemical and crystallographic methods. The P456 crystal structure provided a detailed view of how coaxial stacking stabilizes the packing of RNA helices into tertiary structures. In the self-splicing group II intron from Oceanobacillus iheyensis, the IA and IB stems coaxially stack and contribute to the relative orientation of the constituent helices of a five-way junction. This orientation facilitates proper folding of the active site of the functional ribozyme. The ribosome contains numerous examples of coaxial stacking, including stacked segments as long as 70 bp.

Two common motifs involving coaxial stacking are kissing loops and pseudoknots. In kissing loop interactions, the single-stranded loop regions of two hairpins interact through base pairing, forming a composite, coaxially stacked helix. Notably, this structure allows all of the nucleotides in each loop to participate in base-pairing and stacking interactions. This motif was visualized and studied using NMR analysis by Lee and Crothers. The pseudoknot motif occurs when a single stranded region of a hairpin loop basepairs with an upstream or downstream sequence within the same RNA strand. The two resulting duplex regions often stack upon one another, forming a stable coaxially stacked composite helix. One example of a pseudoknot motif is the highly stable Hepatitis Delta virus ribozyme, in which the backbone shows an overall double pseudoknot topology.

An effect similar to coaxial stacking has been observed in rationally designed DNA structures. DNA origami structures contain a large number of double helixes with exposed blunt ends. These structures were observed to stick together along the edges that contained these exposed blunt ends, due to the hydrophobic stacking interactions.

Other motifs

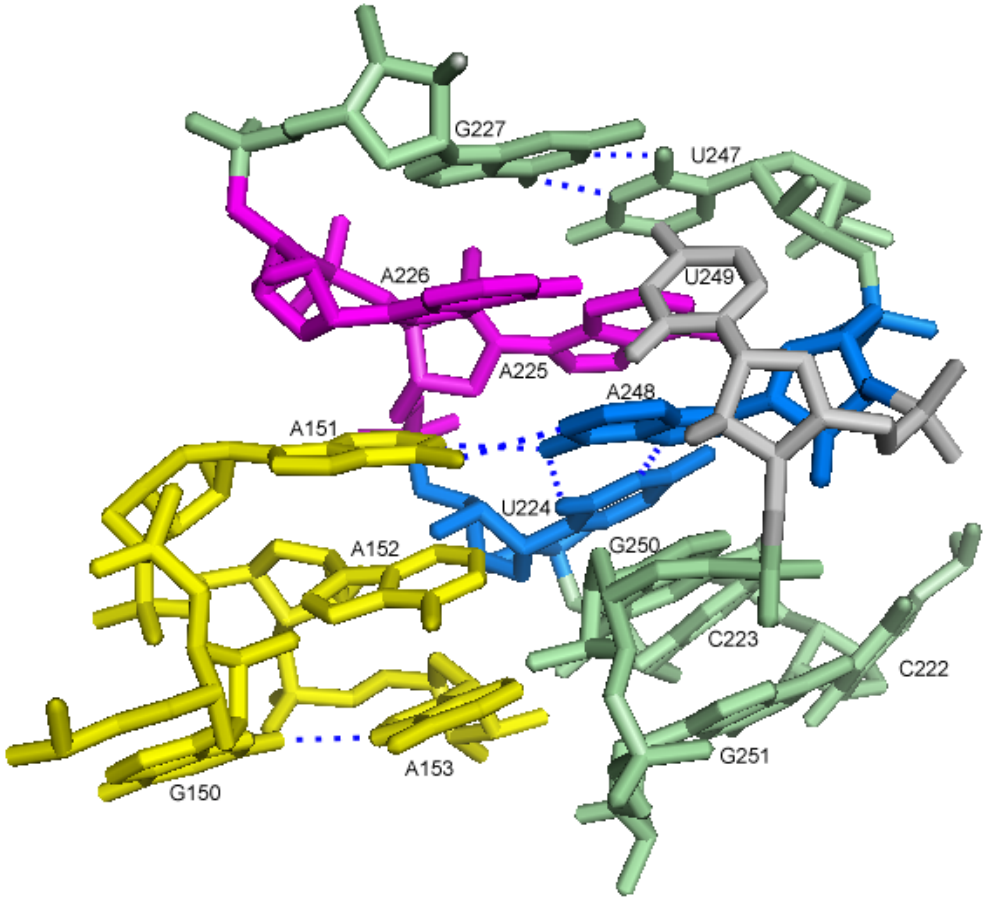
Tetraloop-receptor interactions



Stick representation of a GAAA tetraloop - an example from the GNRA tetraloop family.

Tetraloop-receptor interactions combine base-pairing and stacking interactions between the loop nucleotides of a tetraloop motif and a receptor motif located within an RNA duplex, creating a tertiary contact that stabilizes the global tertiary fold of an RNA molecule. Tetraloops are also possible structures in DNA duplexes.

Stem-loops can vary greatly in size and sequence, but tetraloops of four nucleotides are very common and they usually belong to one of three categories, based on sequence. These three families are the CUYG, UNCG and GNRA tetraloops. In each of these tetraloop families, the second and third nucleotides form a turn in the RNA strand and a base-pair between the first and fourth nucleotides stabilizes the stemloop structure. It has been determined, in general, that the stability of the tetraloop depends on the composition of bases within the loop and on the composition of this "closing base pair". The GNRA family of tetraloops is the most commonly observed within Tetraloop-receptor interactions.



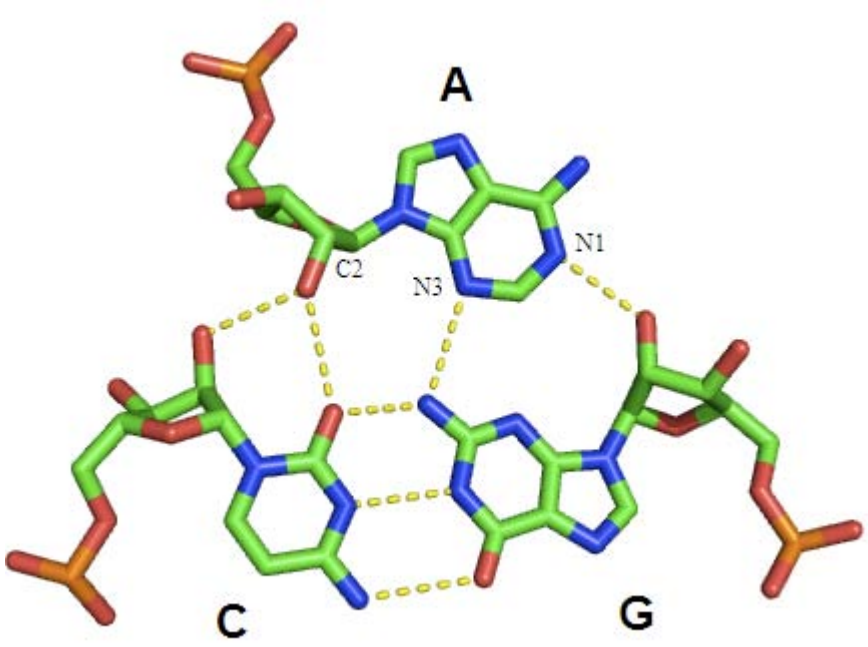
GAAA Tetraloop and Receptor: Stick representation of tetraloop (yellow) and its receptor, showing both Watson-Crick and Hoogsteen base-pairing.

"Tetraloop receptor motifs" are long-range tertiary interactions consisting of hydrogen bonding between the bases in the tetraloop to stemloop sequences in distal sections of the secondary RNA structure. In addition to hydrogen bonding, stacking interactions are an important component of these tertiary interactions. For example, in GNRA-tetraloop interactions, the second nucleotide of the tetraloop stacks directly on an A-platform motif within the receptor. The sequence of the tetraloop and its receptor often covary so that the same type of tertiary contact can be made with different isoforms of the tetraloop and its cognate receptor.

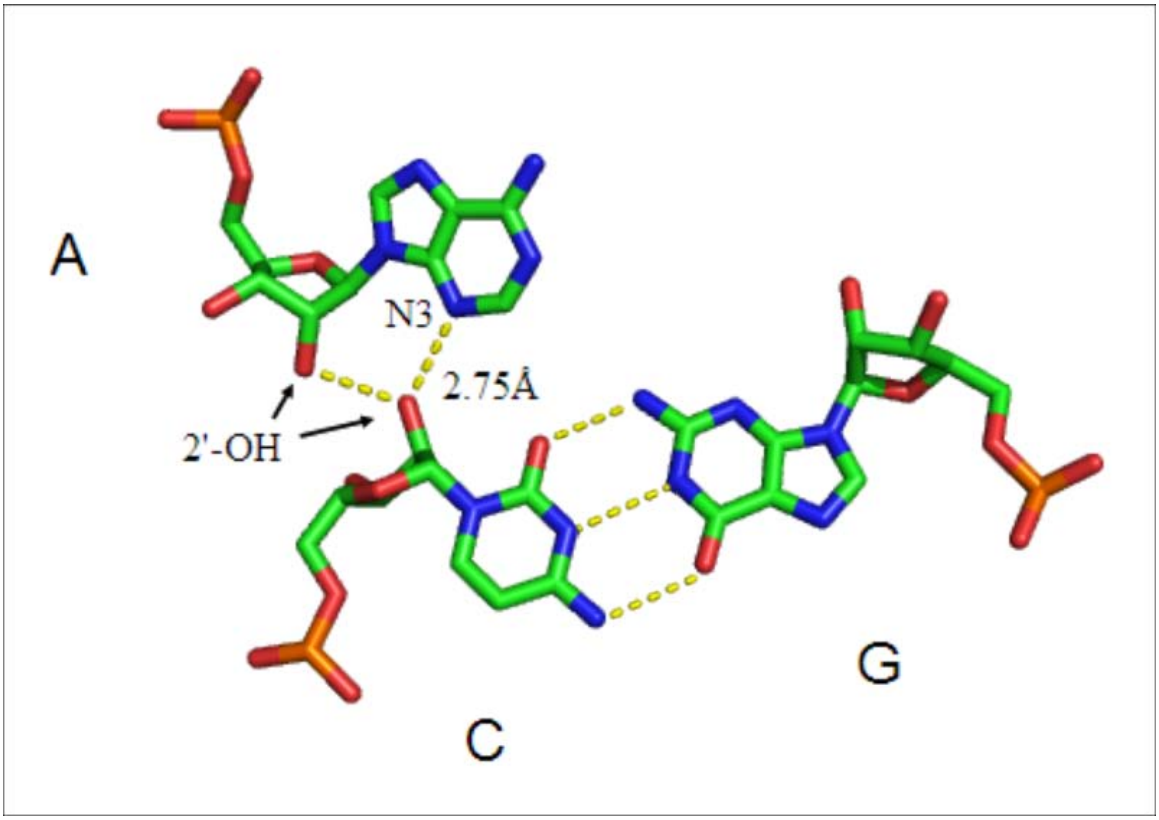
For example, the self-splicing group I intron relies on tetraloop receptor motifs for its structure and function. Specifically, the three adenine residues of the canonical GAAA motif stack on top of the receptor helix and form multiple stabilizing hydrogen bonds with the receptor. The first adenine of the GAAA sequence forms a triple base-pair with the receptor AU bases. The second adenine is stabilized by hydrogen bonds with the same uridine, as well as via its 2'-OH with the recptor and via interactions with the guanine of the GAAA tetraloop. The third adenine forms a triple base pair.

A-minor motif

A-minor Interactions



Type I A-minor interaction: Type I interactions are the most common, strongest A-minor interactions, as they involve numerous hydrogen bonds, and bury the incoming A base in the minor groove.



Type II A-minor interaction: Type II interactions involve the 2'-OH group and N3 of the adenosine. The adenosine interacts with the cytosine's 2'-OH group in the minor groove. The strength of this interaction is on the order of the Type I interaction.

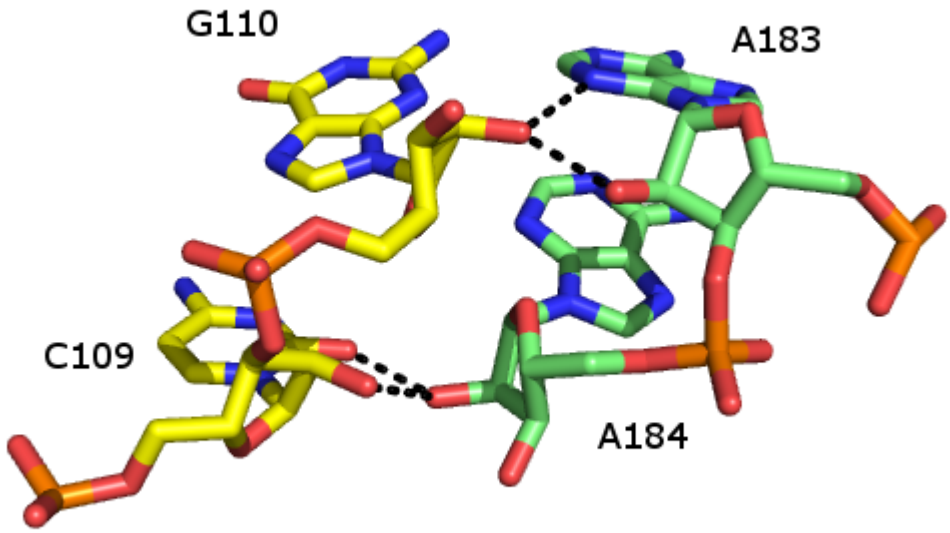
The A-minor motif is a ubiquitous RNA tertiary structural motif. It is formed by the insertion of an unpaired nucleoside into the minor groove of an RNA duplex. As such it is an example of a minor groove triple. Although guanosine, cytosine and uridine can also form minor groove triple interactions, minor groove interactions by adenine are very common. In the case of adenine, the N1-C2-N3 edge of the inserting base forms hydrogen bonds with one or both of the 2'-OH's of the duplex, as well as the bases of the duplex. The host duplex is often a G-C basepair.

A-minor motifs have been separated into four classes, types 0 to III, based upon the position of the inserting base relative to the two 2'-OH's of the Watson-Crick base pair. In type I and II A-minor motifs, N3 of adenine is inserted deeply within the minor groove of the duplex, and there is good shape complementarity with the base pair. Unlike types 0 and III, type I and II interactions are specific for adenine due to hydrogen bonding interactions. In the type III interaction, both the O2' and N3 of the inserting base are associated less closely with the minor groove of the duplex. Type 0 and III motifs are weaker and non-specific because they are mediated by interactions with a single 2'-OH.

The A-minor motif is among the most common RNA structural motifs in the ribosome, where it contributes to the binding of tRNA to the 23S subunit. They most often stabilize RNA duplex interactions in loops and helices, such as in the core of group II introns.

An interesting example of A-minor is its role in anticodon recognition. The ribosome must discriminate between correct and incorrect codon-anticodon pairs. It does so, in part, through the insertion of adenine bases into the minor groove. Incorrect codon-anticodon pairs will present distorted helical geometry, which will prevent the A-minor interaction from stabilizing the binding, and increase the dissociation rate of the incorrect tRNA.

Ribose zipper



Ribose Zippers: View of a canonical ribose zipper between two RNA backbones.

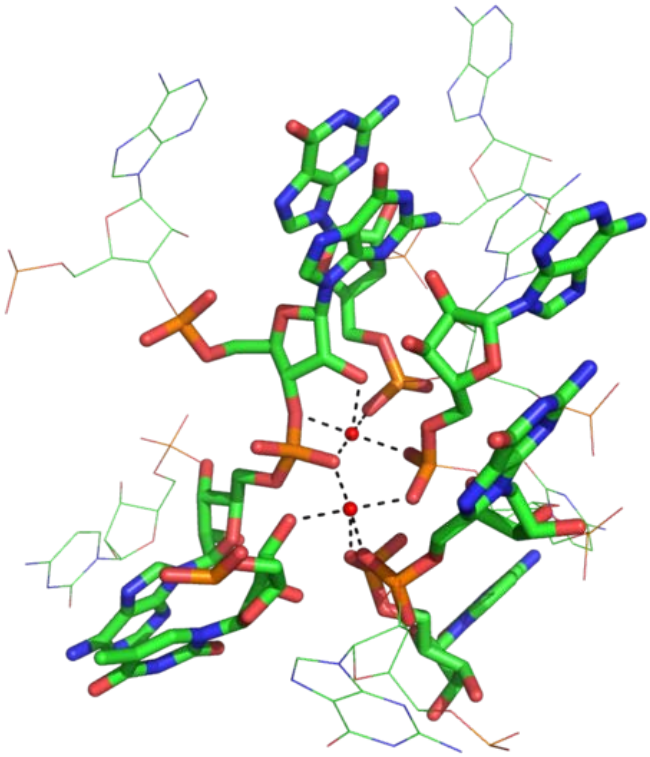
The ribose zipper is an RNA tertiary structural element in which two RNA chains are held together by hydrogen bonding interactions involving the 2'OH of ribose sugars on different strands. The 2'OH can behave as both hydrogen bond donor and acceptor, which allows formation of bifurcated hydrogen bonds with another 2' OH.

Numerous forms of ribose zipper have been reported, but a common type involves four hydrogen bonds between 2'-OH groups of two adjacent sugars. Ribose zippers commonly occur in arrays that stabilize interactions between separate RNA strands. Ribose zippers are often observed as Stem-loop interactions with very low sequence specificity. However, in the small and large ribosomal subunits, there exists a propensity for ribose

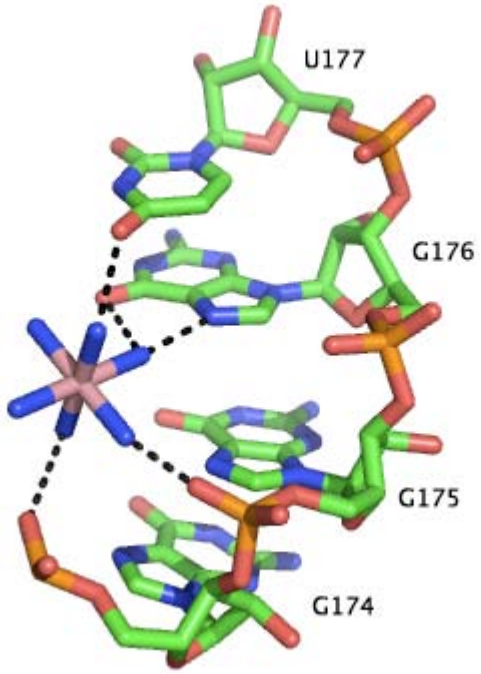
zippers of the CC/AA sequence- two cytosines on the first chain paired to two adenines on the second chain.

Role of metal ions

Metal Ion Binding in the Group I Intron



PDB rendering of Group I intron inner sphere magnesium coordination. The two red balls indicate magnesium ions and dashed lines coming from the ions indicate coordination with the respective groups on nucleotides. The color-coding scheme is as follows: green=carbon, orange=phosphate, pink=oxygen, blue=nitrogen.



PDB rendering of Group 1 intron P5c binding pocket demonstrating outer sphere coordination. Here, the six amines of osmium hexamine(III) fulfill the role generally served by water molecules and mediate the ion's interaction with the major groove. Coordination via hydrogen bonds is indicated by dashed lines and osmium is rendered in pink, all other colors are as above.

Functional RNAs are often folded, stable molecules with three-dimensional shapes rather than floppy, linear strands. Cations are essential for thermodynamic stabilization of RNA tertiary structures. Metal cations that bind RNA can be monovalent, divalent or trivalent. Potassium (K^+) is a common monovalent ion that binds RNA. A common divalent ion that binds RNA is magnesium (Mg^{2+}). Other ions including sodium (Na^+), calcium (Ca^{2+}) and manganese (Mn^{2+}) have been found to bind RNA *in vivo* and *in vitro*. Multivalent organic cations such as spermidine or spermine are also found in cells and these make important contributions to RNA folding. Trivalent ions such as cobalt hexamine or lanthanide ions such as terbium (Tb^{3+}) are useful experimental tools for studying metal binding to RNA.

A metal ion can interact with RNA in multiple ways. An ion can associate diffusely with the RNA backbone, shielding otherwise unfavorable electrostatic interactions. This charge screening is often fulfilled by monovalent ions. Site-bound ions stabilize specific elements of RNA tertiary structure. Site-bound interactions can be further subdivided into two categories depending on whether water mediates the metal binding. "Outer sphere" interactions are mediated by water molecules that surround the metal ion. For example, magnesium hexahydrate interacts with and stabilizes specific RNA tertiary structure motifs via interactions with guanosine in the major groove. Conversely, "inner sphere"

interactions are directly mediated by the metal ion. RNA often folds in multiple stages and these steps can be stabilized by different types of cations. In the early stages, RNA forms secondary structures stabilized through the binding of monovalent cations, divalent cations and polyanionic amines in order to neutralize the polyanionic backbone. The later stages of this process involve the formation of RNA tertiary structure, which is stabilized almost largely through the binding of divalent ions such as magnesium with possible contributions from potassium binding.

Metal-binding sites are often localized in the deep and narrow major groove of the RNA duplex, coordinating to the Hoogsteen edges of purines. In particular, metal cations stabilize sites of backbone twisting where tight packing of phosphates results in a region of dense negative charge. There are several metal ion-binding motifs in RNA duplexes that have been identified in crystal structures. For instance, in the P4-P6 domain of the *Tetrahymena thermophila* group I intron, several ion-binding sites consist of tandem G-U wobble pairs and tandem G-A mismatches, in which divalent cations interact with the Hoogsteen edge of guanosine via O6 and N7. Another ion-binding motif in the *Tetrahymena* group I intron is the A-A platform motif, in which consecutive adenosines in the same strand of RNA form a non-canonical pseudobase pair. Unlike the tandem G-U motif, the A-A platform motif binds preferentially to monovalent cations. In many of these motifs, absence of the monovalent or divalent cations results in either greater flexibility or loss of tertiary structure.

Divalent metal ions, especially magnesium, have been found to be important for the structure of DNA junctions such as the Holliday junction intermediate in genetic recombination. The magnesium ion shields the negatively-charged phosphate groups in the junction and allows them to be positioned closer together, allowing a stacked conformation rather than an unstacked conformation. Magnesium is vital in stabilizing these kinds of junctions in artificially designed structures used in DNA nanotechnology, such as the double crossover motif.

History

The earliest work in RNA structural biology coincided, more or less, with the work being done on DNA in the early 1950s. In their seminal 1953 paper, Watson and Crick suggested that van der Waals crowding by the 2`OH group of ribose would preclude RNA from adopting a double helical structure identical to the model they proposed - what we now know as B-form DNA. This provoked questions about the three dimensional structure of RNA: could this molecule form some type of helical structure, and if so, how?

In the mid 1960's, the role of tRNA in protein synthesis was being intensively studied. In 1965, Holley *et al.* purified and sequenced the first tRNA molecule, initially proposing that it adopted a cloverleaf structure, based largely on the ability of certain regions of the molecule to form stem loop structures. The isolation of tRNA proved to be the first major windfall in RNA structural biology. In 1971, Kim *et al.* achieved another breakthrough,

producing crystals of yeast $tRNA^{PHE}$ that diffracted to 2-3 Ångström resolutions by using spermine, a naturally occurring polyamine, which bound to and stabilized the tRNA.

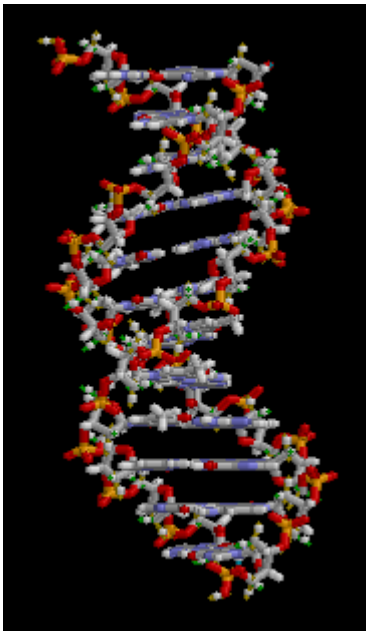
For a considerable time following the first tRNA structures, the field of RNA structure did not dramatically advance. The ability to study an RNA structure depended upon the potential to isolate the RNA target. This proved limiting to the field for many years, in part owing to the fact that other known targets - i.e. the ribosome - were significantly more difficult to isolate and crystallize. As such, for some twenty years following the original publication of the $tRNA^{PHE}$ structure, the structures of only a handful of other RNA targets were solved, with almost all of these belonging to the transfer RNA family.

This unfortunate lack of scope would eventually be overcome largely because of two major advancements in nucleic acid research: the identification of ribozymes, and the ability to produce them via *in vitro* transcription. Subsequent to Tom Cech's publication implicating the *Tetrahymena* group I intron as an autocatalytic ribozyme, and Sidney Altman's report of catalysis by ribonuclease P RNA, several other catalytic RNAs were identified in the late 1980s, including the hammerhead ribozyme. In 1994, McKay *et al.* published the structure of a 'hammerhead RNA-DNA ribozyme-inhibitor complex' at 2.6 Ångström resolution, in which the autocatalytic activity of the ribozyme was disrupted via binding to a DNA substrate. In addition to the advances being made in global structure determination via crystallography, the early 1990s also saw the implementation of NMR as a powerful technique in RNA structural biology. Investigations such as this enabled a more precise characterization of the base pairing and base stacking interactions which stabilized the global folds of large RNA molecules.

The resurgence of RNA structural biology in the mid-1990s has caused a veritable explosion in the field of nucleic acid structural research. Since the publication of the hammerhead and P_{4-6} structures, numerous major contributions to the field have been made. Some of the most noteworthy examples include the structures of the Group I and Group II introns, and the Ribosome. It should be noted that the first three structures were produced using *in vitro* transcription, and that NMR has played a role in investigating partial components of all four structures - testaments to the indispensability of both techniques for RNA research. Most recently, the 2009 Nobel Prize in Chemistry was awarded to Ada Yonath, Venkatraman Ramakrishnan, and Thomas Steitz for their structural work on the ribosome, demonstrating the prominent role RNA structural biology has taken in modern molecular biology.

Chapter 4

DNA



The structure of part of a DNA double helix

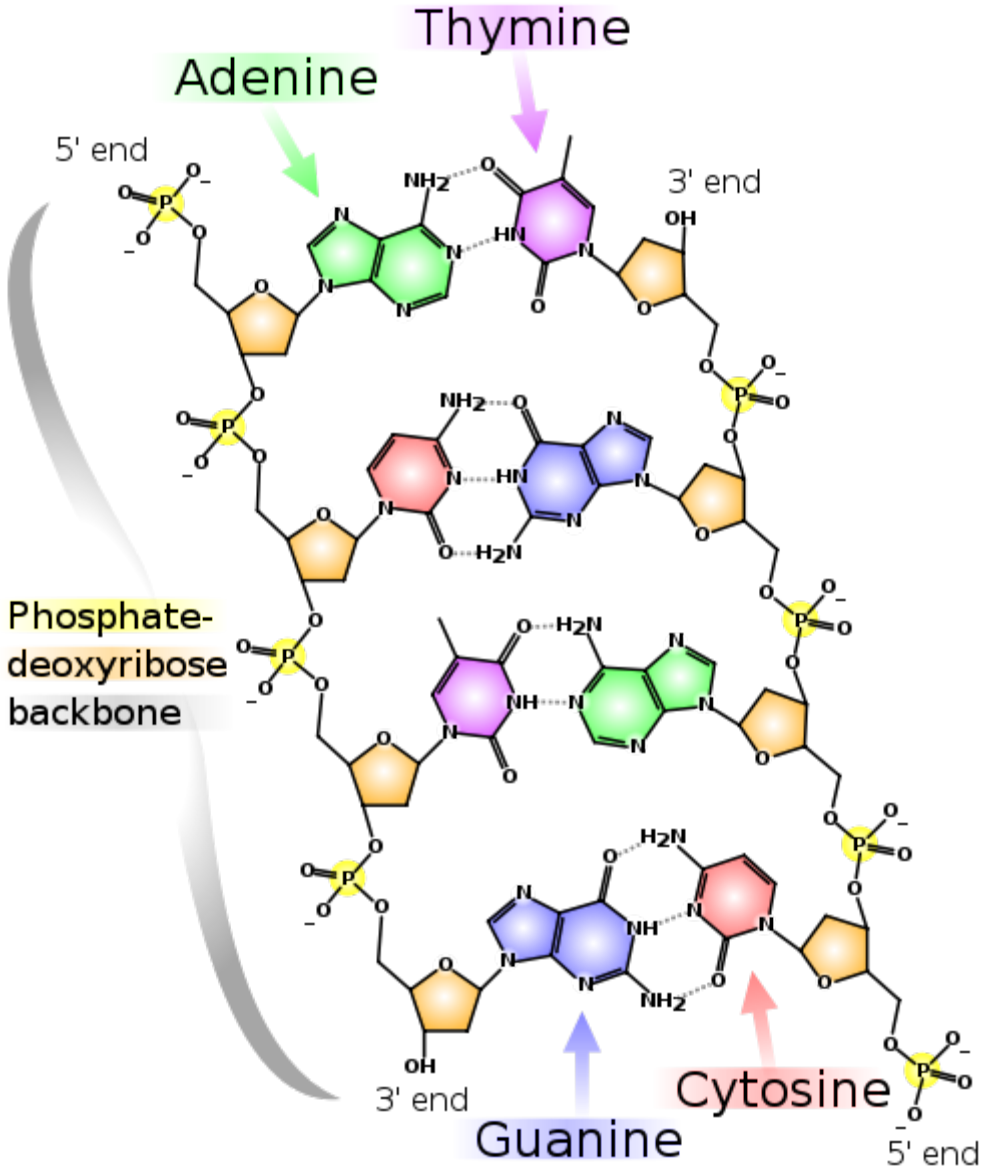
Deoxyribonucleic acid is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms (with the exception of RNA viruses). The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints,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



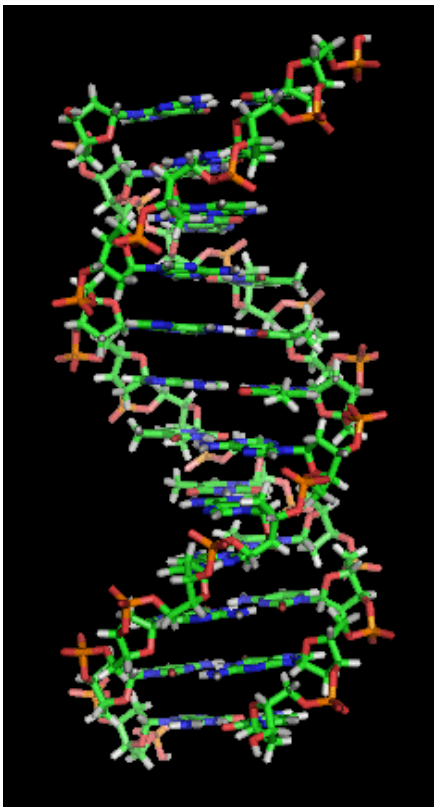
Chemical structure of DNA. Hydrogen bonds shown as dotted lines.

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.

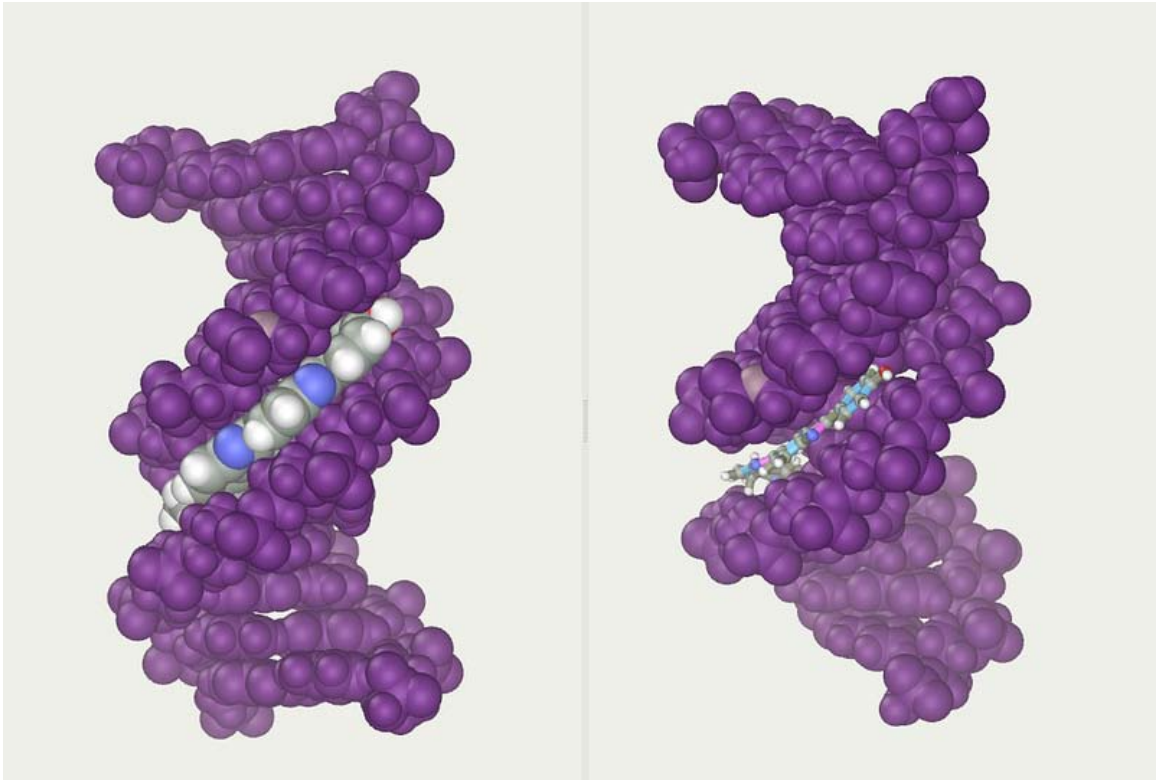
The backbone of the DNA strand is made from alternating phosphate and sugar residues. The sugar in DNA is 2-deoxyribose, which is a pentose (five-carbon) sugar. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar 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 primarily by two forces: hydrogen bonds between nucleotides and base-stacking interactions among the aromatic bases. In the aqueous environment of the cell, the conjugated π bonds of nucleotide bases align perpendicular to the axis of the DNA molecule, minimizing their interaction with the solvation shell and therefore, the Gibbs free energy. The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). These four bases are attached to the sugar/phosphate to form the complete nucleotide, as shown for adenosine monophosphate.

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.



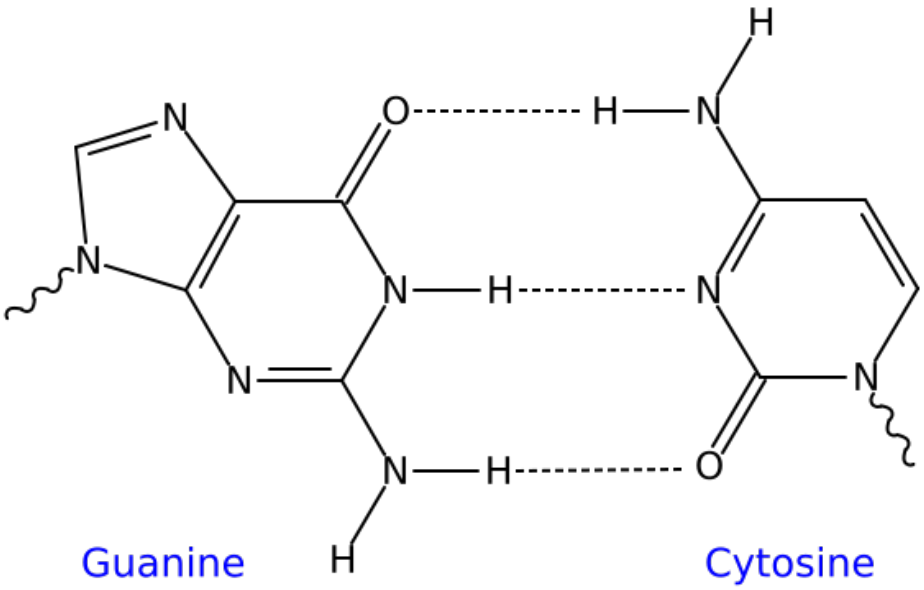
Major and minor grooves of DNA. Minor groove is a binding site for the dye Hoechst 33258.

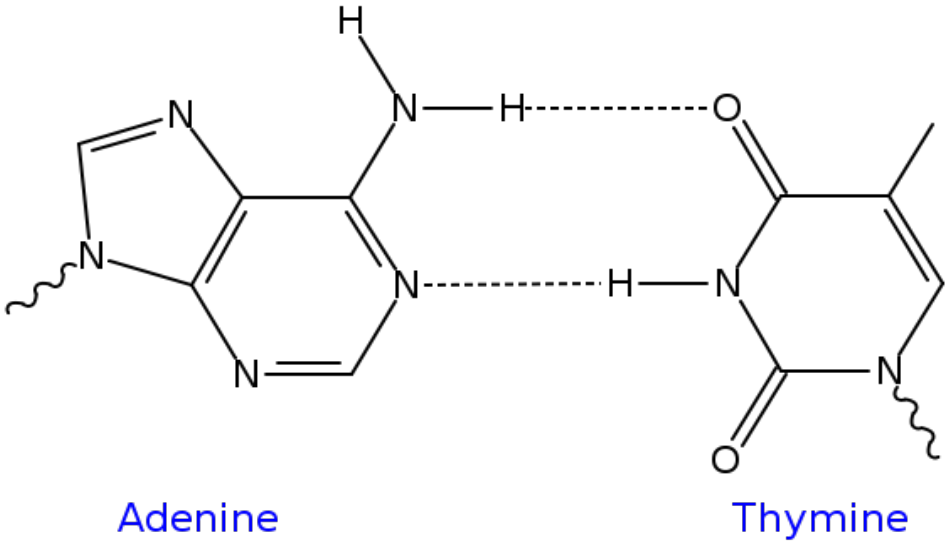
Grooves

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, 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, due to the added stability of an additional hydrogen bond.

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

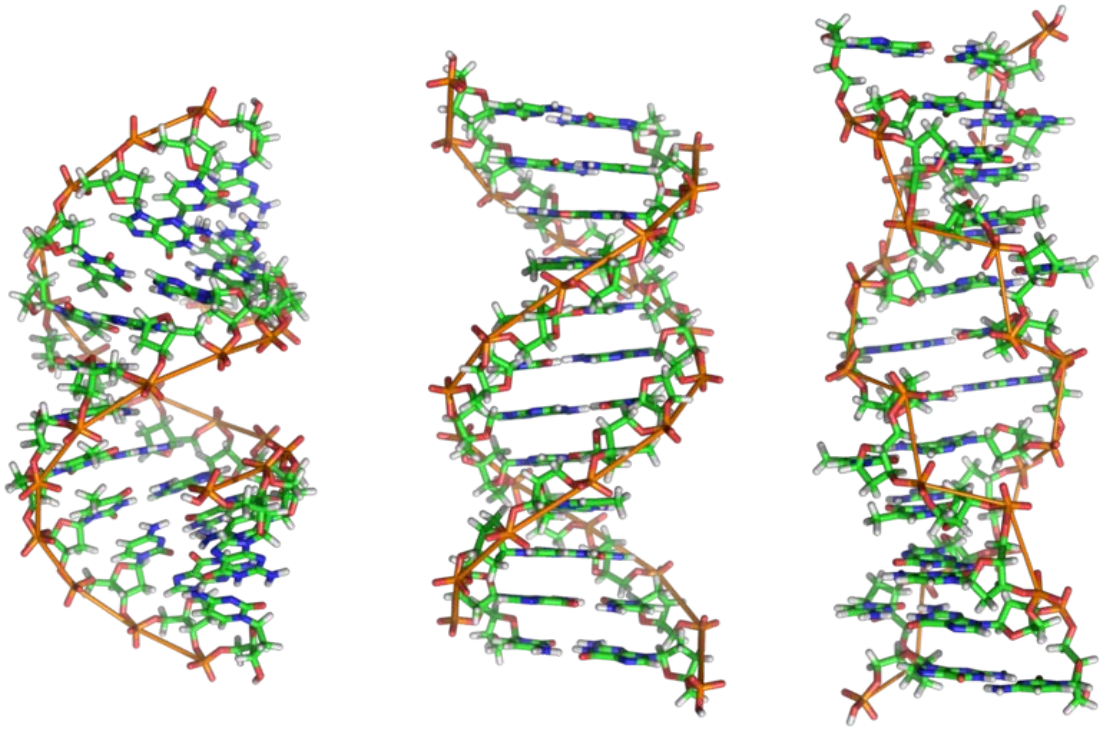
A DNA sequence is called "sense" if its sequence is the same as that of a messenger RNA copy that is translated into protein. The sequence on the opposite strand is called the "antisense" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands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

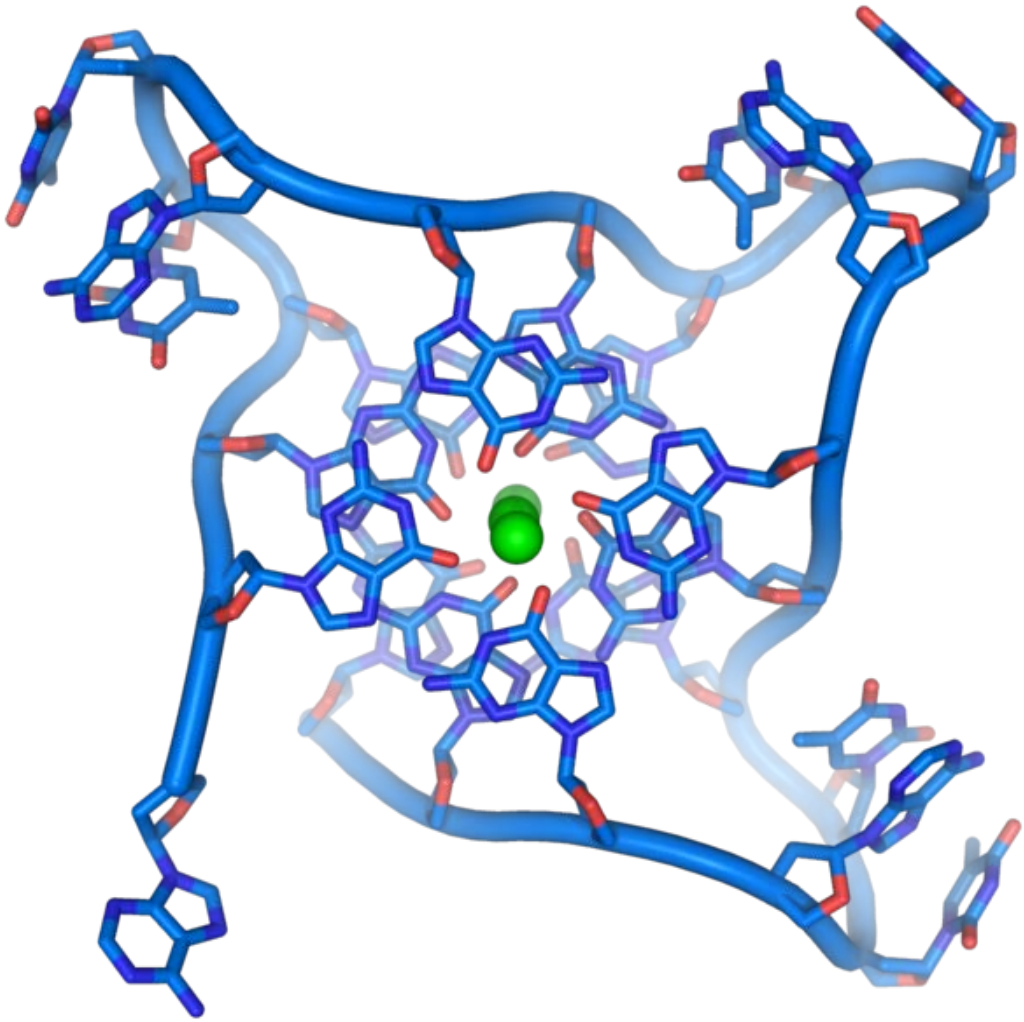
For a number of years exobiologists have proposed the existence of a shadow biosphere, a postulated microbial biosphere of Earth that uses radically different biochemical and molecular processes than currently known life. One of the proposals was the existence of lifeforms that use arsenic instead of 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

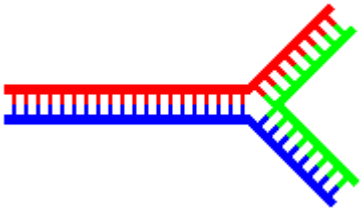
At the ends of the linear chromosomes are specialized regions of DNA called telomeres. The main function of these regions is to allow the cell to replicate chromosome ends using the enzyme telomerase, as the enzymes that normally replicate DNA cannot copy the extreme 3′ ends of chromosomes. These specialized 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.



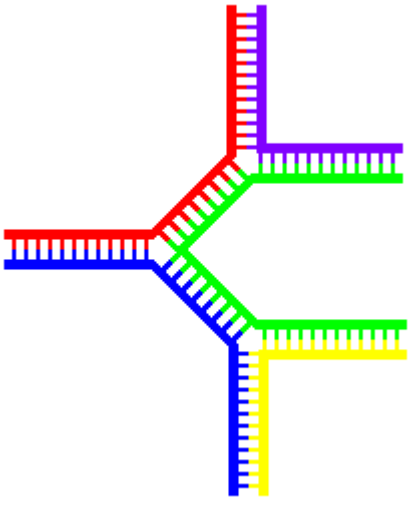
DNA quadruplex formed by telomere repeats. The looped conformation of the DNA backbone is very different from the typical DNA helix.

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.

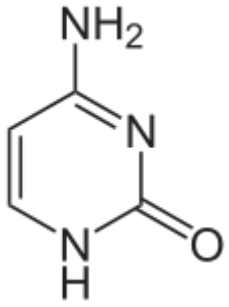
Branched DNA

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.

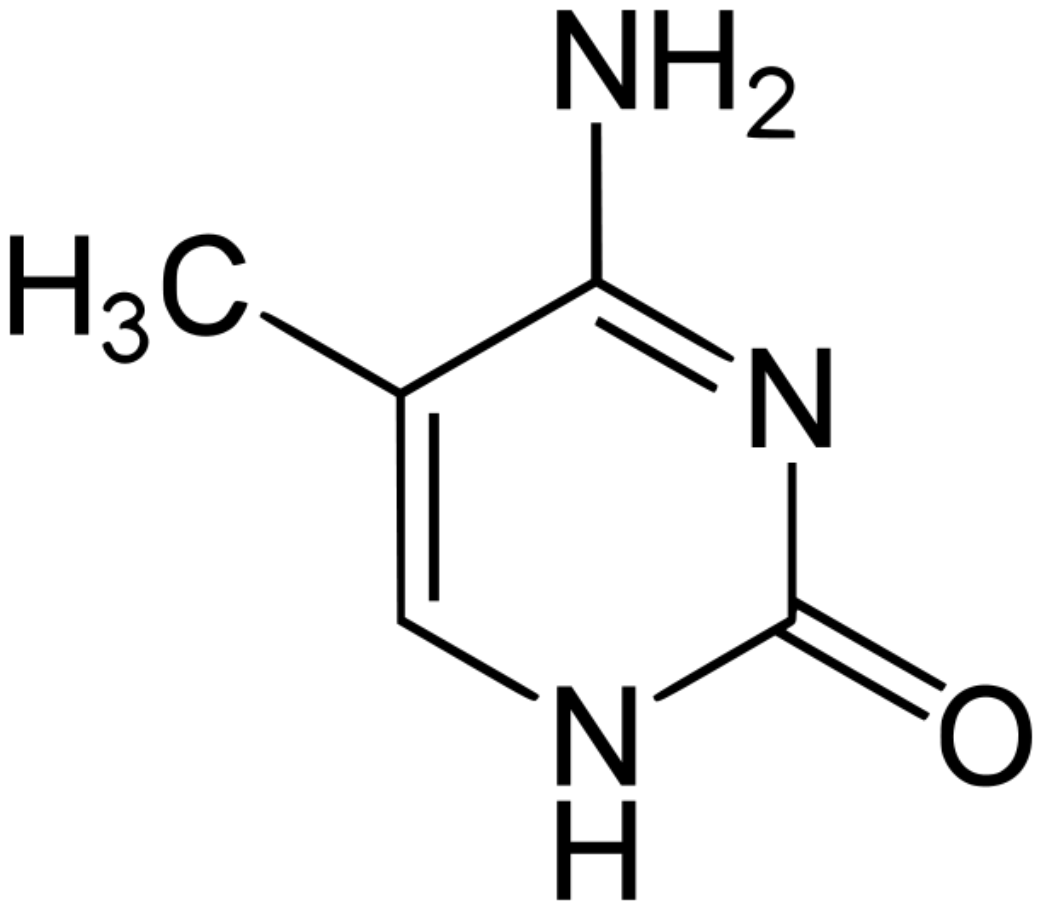
Vibration

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

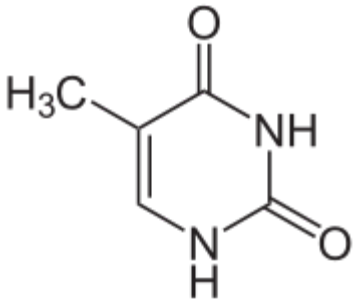
Chemical modifications



cytosine



5-methylcytosine



thymine

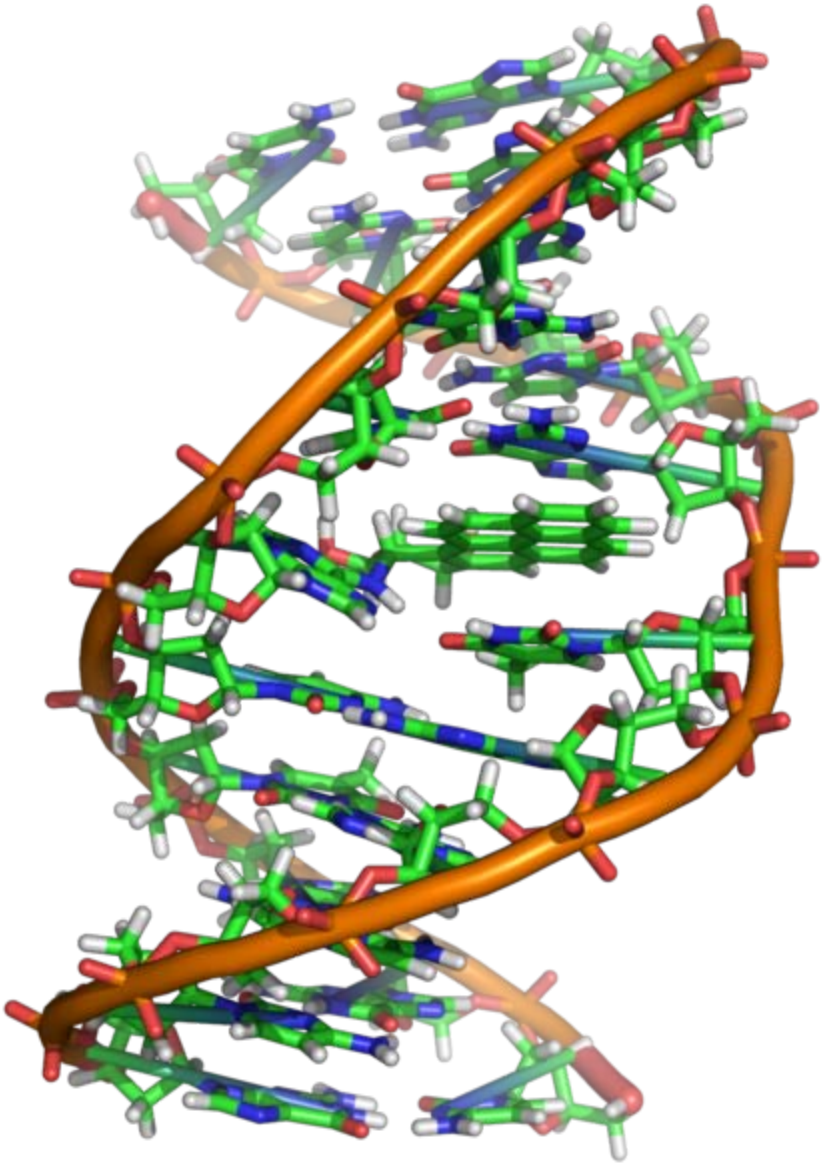
Structure of cytosine with and without the 5-methyl group. Deamination converts 5-methylcytosine into thymine.

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

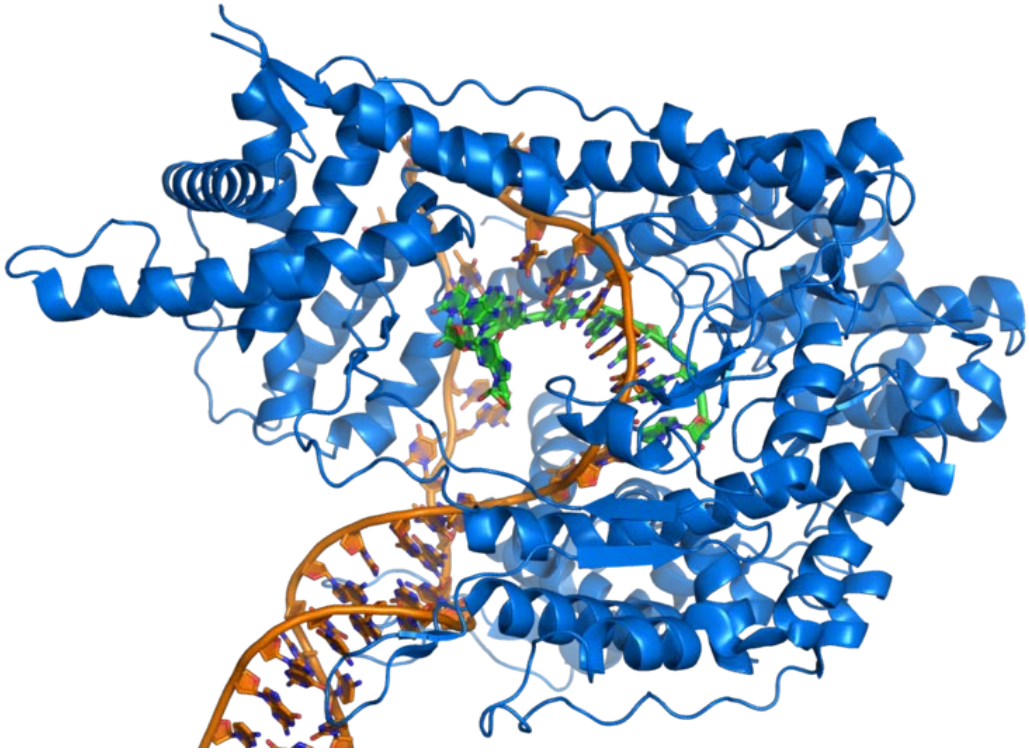
DNA usually occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The set of chromosomes in a cell makes up its genome; the human genome has approximately 3 billion base pairs of DNA arranged into 46 chromosomes. The information carried by DNA is held in the sequence of 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.



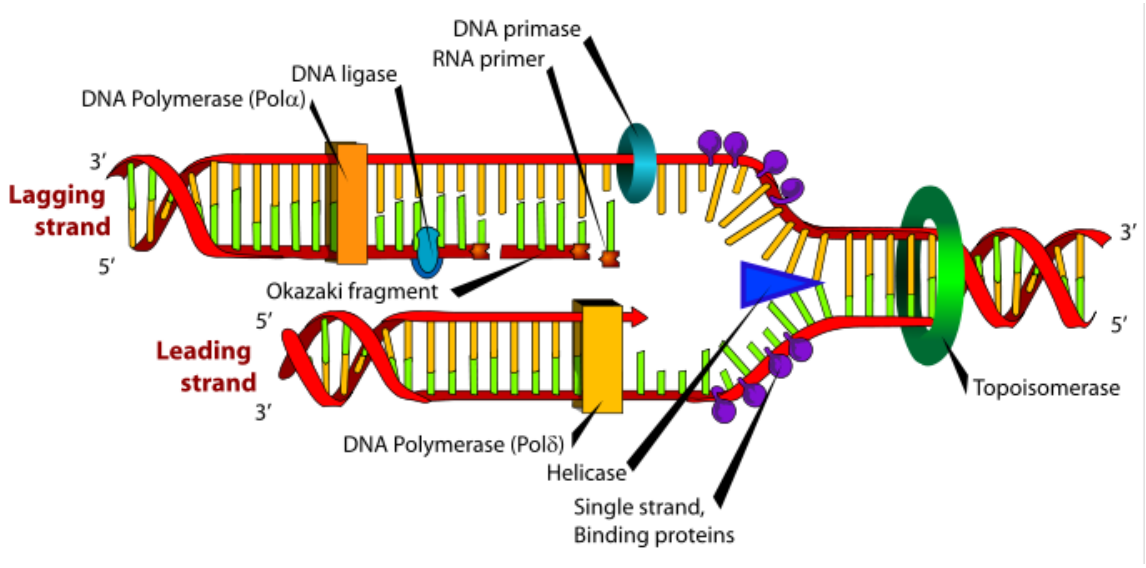
T7 RNA polymerase (blue) producing a mRNA (green) from a DNA template (orange).

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

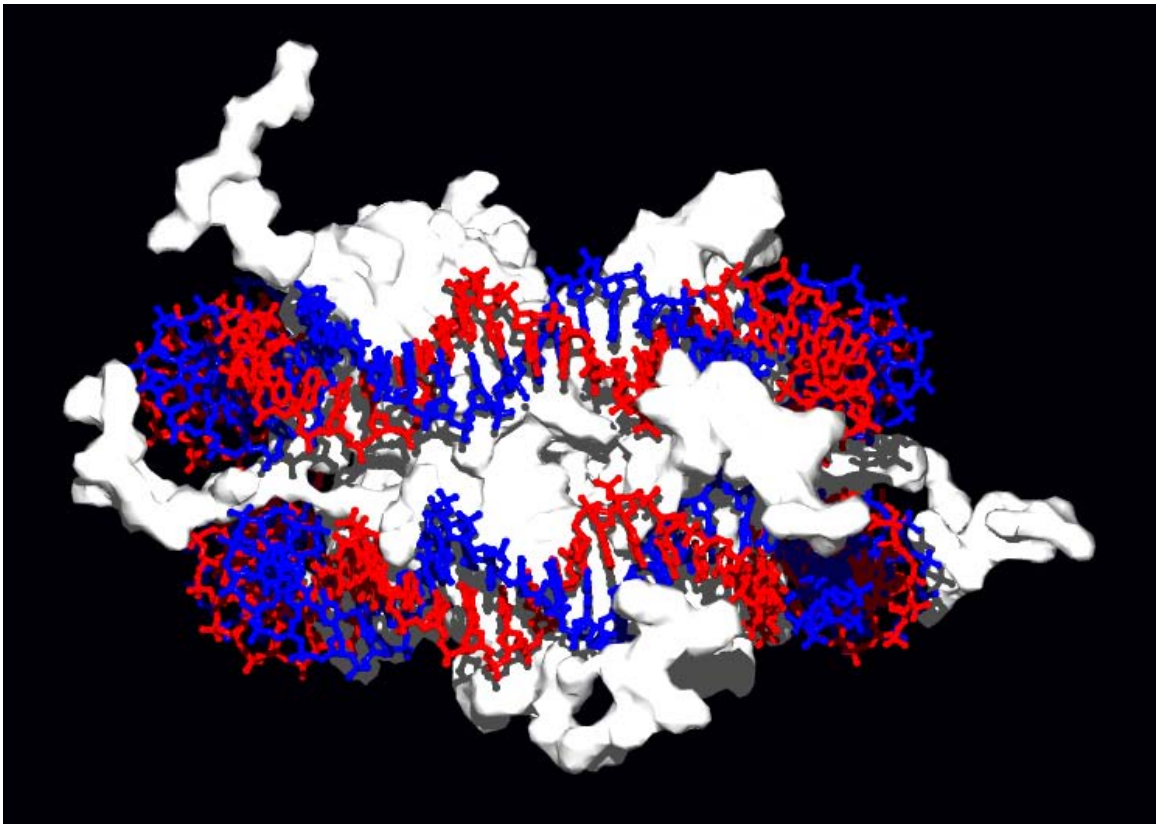
Cell division is essential for an organism to grow, but, when a cell divides, it must replicate the DNA in its genome so that the two daughter cells have the same genetic information as their parent. The double-stranded structure of DNA provides a simple mechanism for DNA replication. Here, the two strands are separated and then each strand's complementary DNA sequence is recreated by an enzyme called DNA

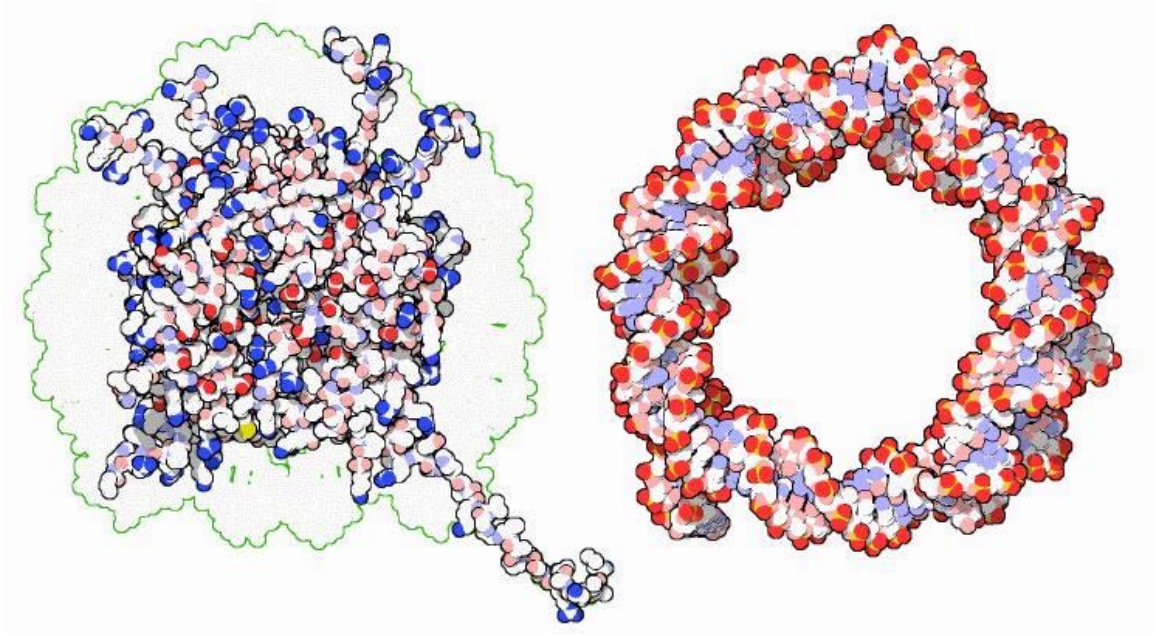
polymerase. This enzyme makes the complementary strand by finding the correct base through complementary base pairing, and bonding it onto the original strand. As DNA polymerases can only extend a DNA strand in a 5′ to 3′ direction, different mechanisms are used to copy the antiparallel strands of the 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

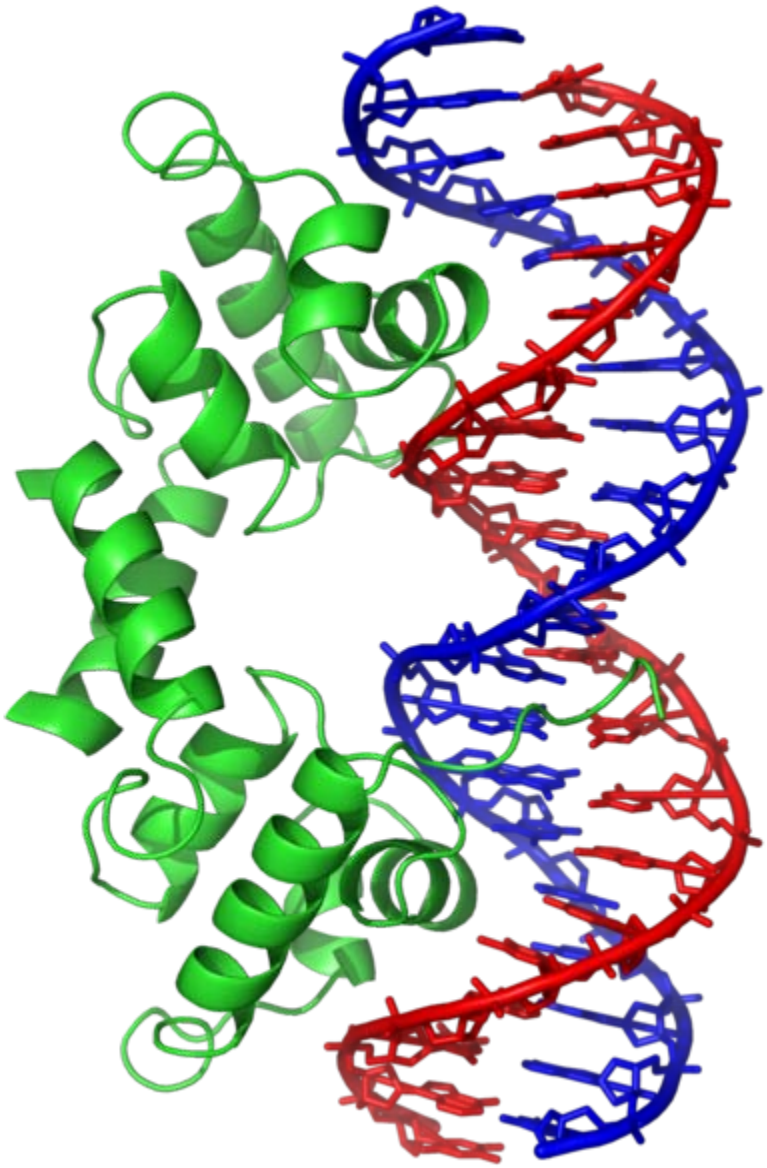




Interaction of DNA with histones (shown in white, top). These proteins' basic amino acids (below left, blue) bind to the acidic phosphate groups on DNA (below right, red).

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.

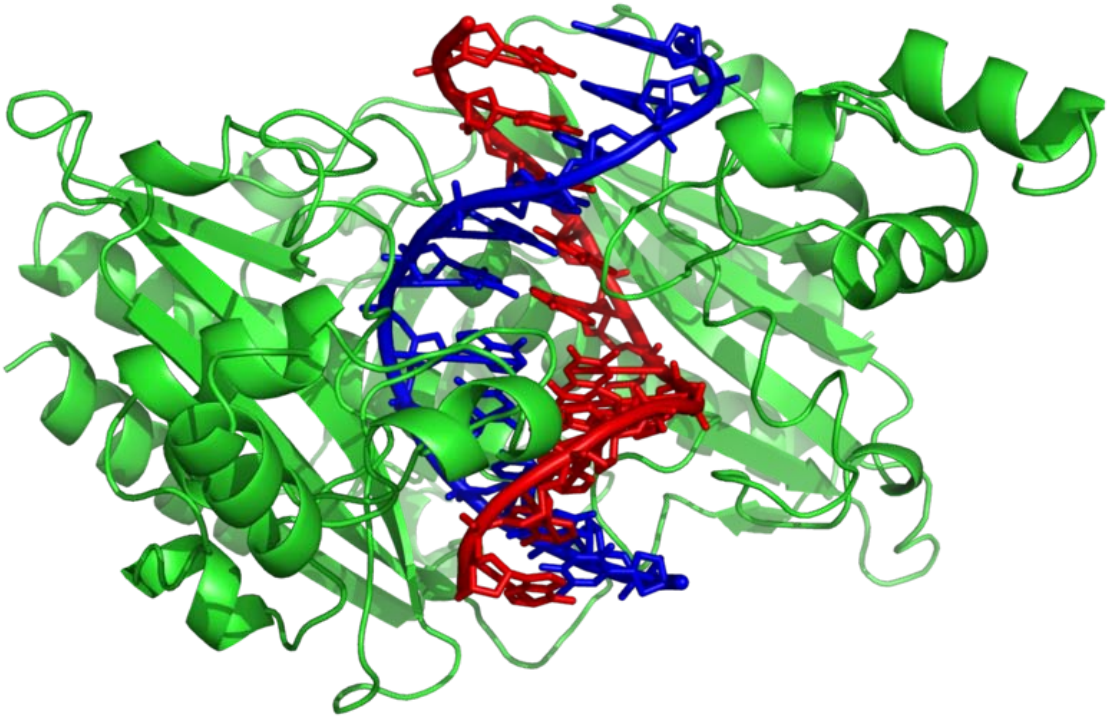
A distinct group of DNA-binding proteins are the DNA-binding proteins that specifically bind single-stranded DNA. In humans, replication protein A is the best-understood member of this family and is used in processes where the double helix is separated, including DNA replication, recombination and DNA 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.

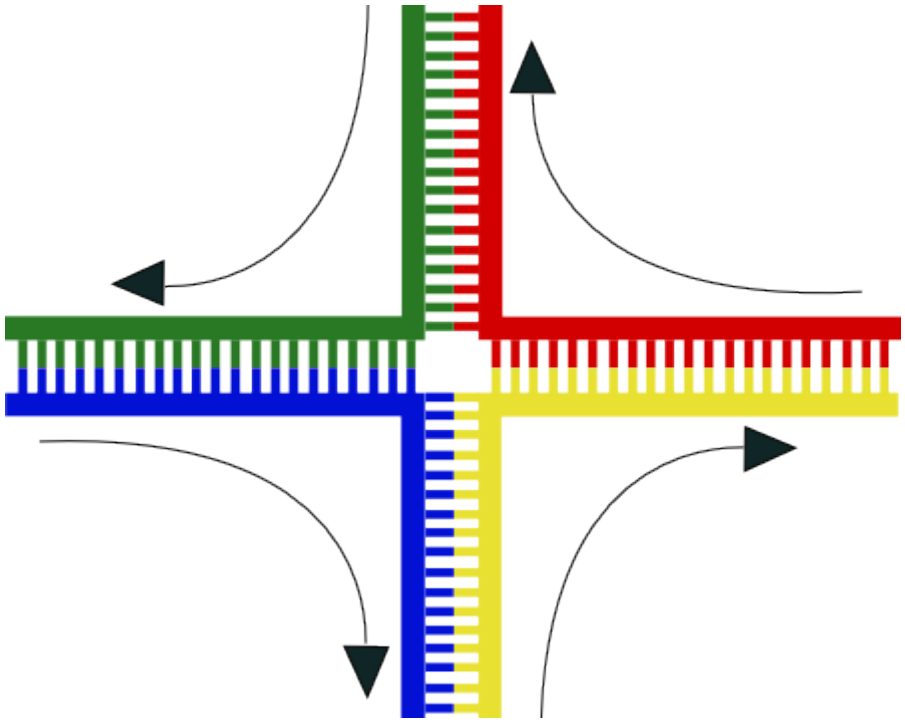
In DNA replication, a DNA-dependent DNA polymerase makes a copy of a DNA sequence. Accuracy is vital in this process, so many of these polymerases have a proofreading activity. Here, the polymerase recognizes the occasional mistakes in the synthesis reaction by the lack of base pairing between the mismatched 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.

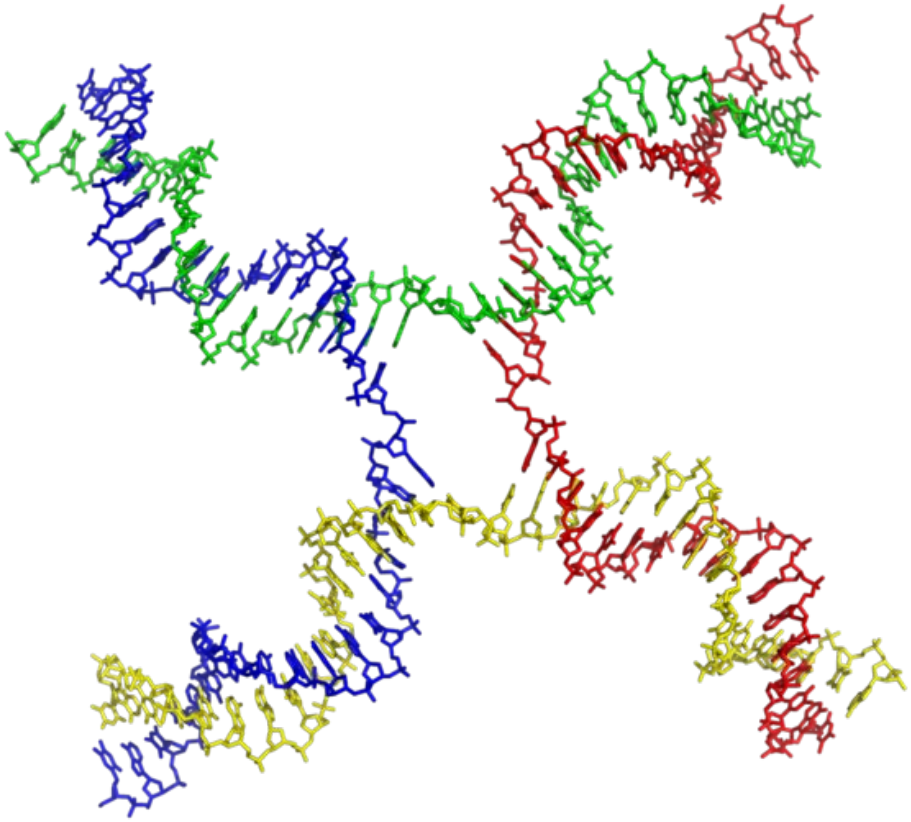
RNA-dependent DNA polymerases are a specialized class of polymerases that copy the sequence of an RNA strand into DNA. They include reverse transcriptase, which is a viral enzyme involved in the infection of cells by retroviruses, and telomerase, which is

required for the replication of telomeres. Telomerase is an unusual polymerase because it contains its own RNA template as part of its structure.

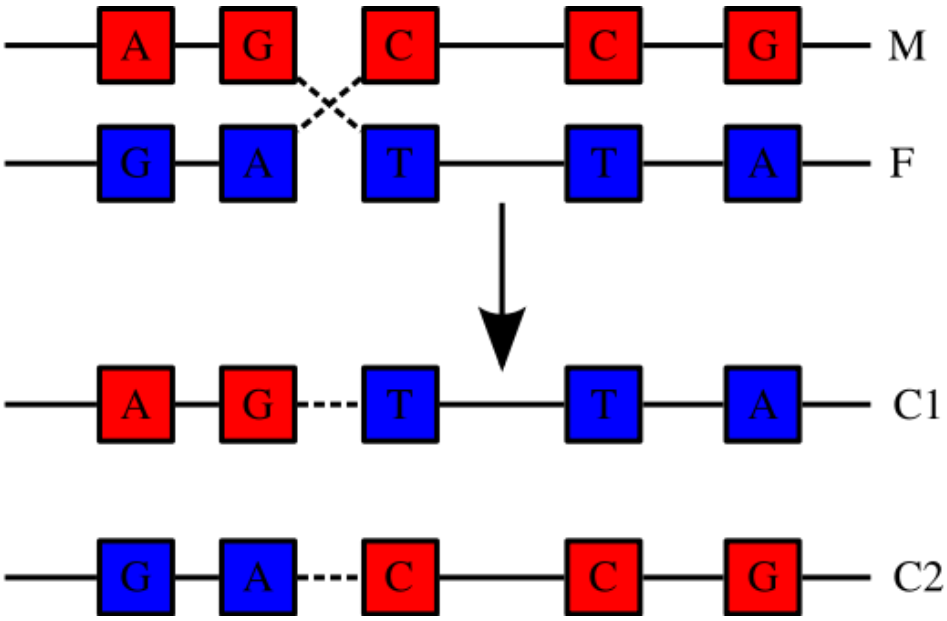
Transcription is carried out by a DNA-dependent RNA polymerase that copies the sequence of a DNA strand into RNA. To begin transcribing a gene, the RNA polymerase binds to a sequence of DNA called a promoter and separates the DNA strands. It then copies the gene sequence into a messenger RNA transcript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

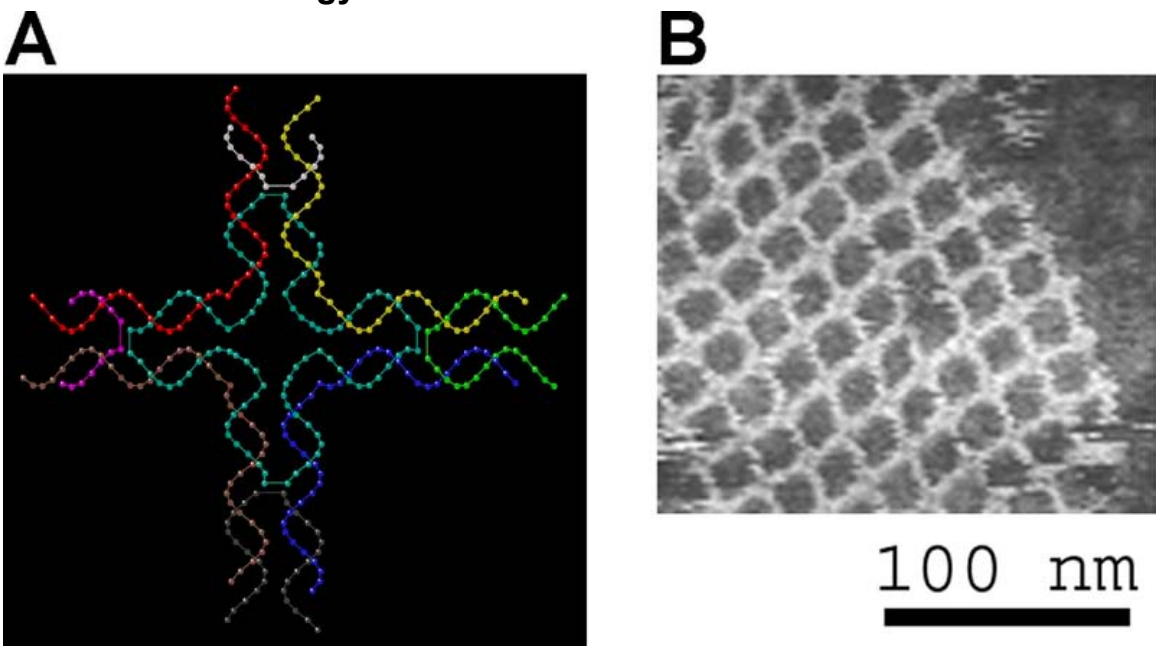
People convicted of certain types of crimes may be required to provide a sample of DNA for a database. This has helped investigators solve old cases where only a DNA sample was obtained from the scene. DNA profiling can also be used to identify victims of mass casualty incidents. On the other hand, many 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. Image from Strong, 2004.

DNA nanotechnology uses the unique molecular recognition properties of DNA and other nucleic acids to create self-assembling branched DNA complexes with useful properties. 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.

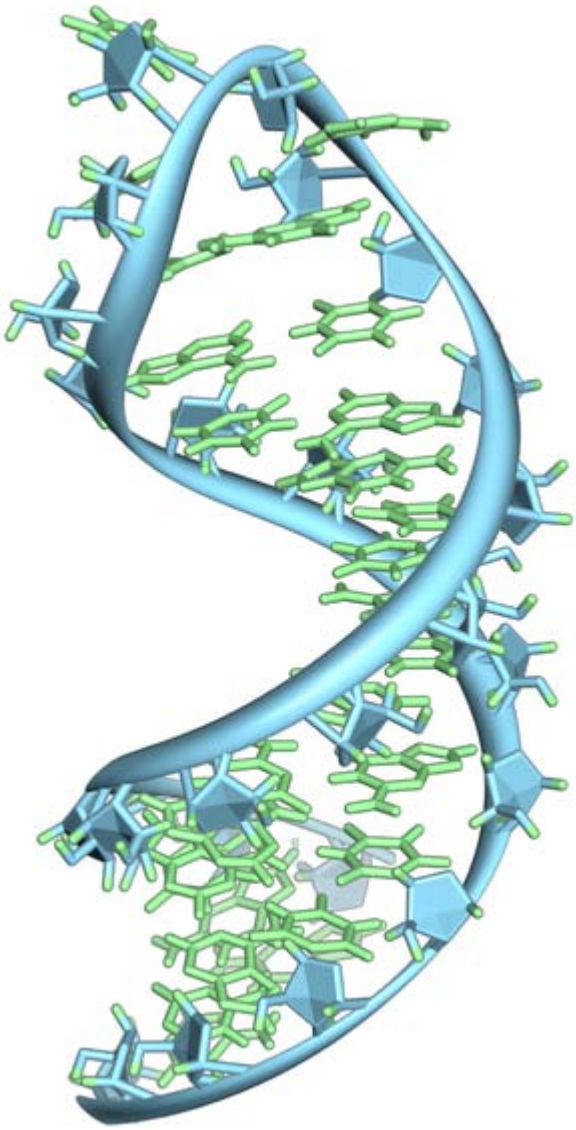
In 1953, James D. Watson and Francis Crick suggested what is now accepted as the first correct double-helix model of DNA structure in the journal *Nature*. Their double-helix, molecular model of DNA was then based on a single X-ray diffraction image (labeled as "Photo 51") taken by Rosalind Franklin 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 5

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.

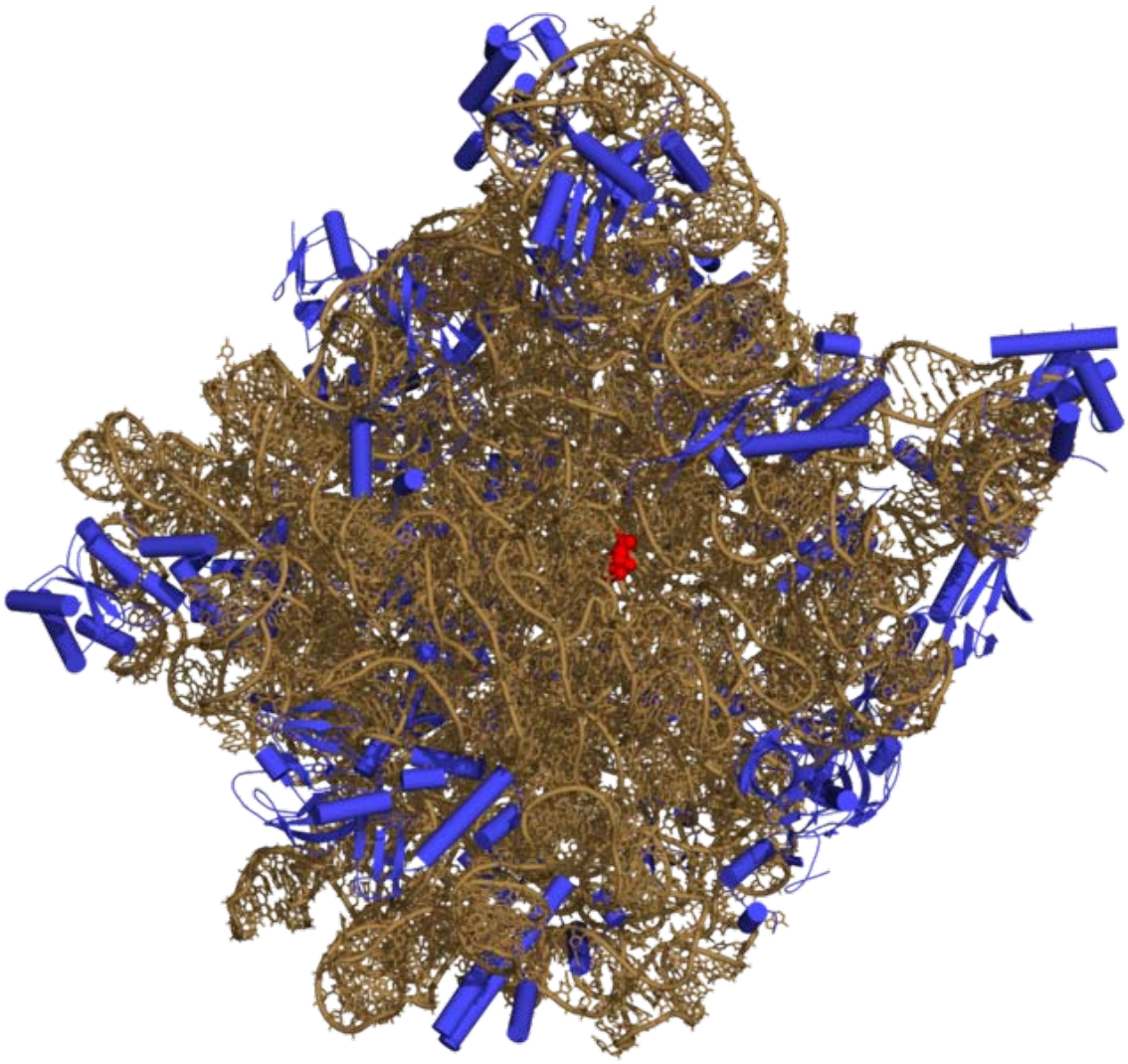
Like DNA, RNA is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase (sometimes called a nitrogenous base), a ribose sugar, and a phosphate group. The sequence of nucleotides allows RNA to encode genetic information. For example, some viruses use RNA instead 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



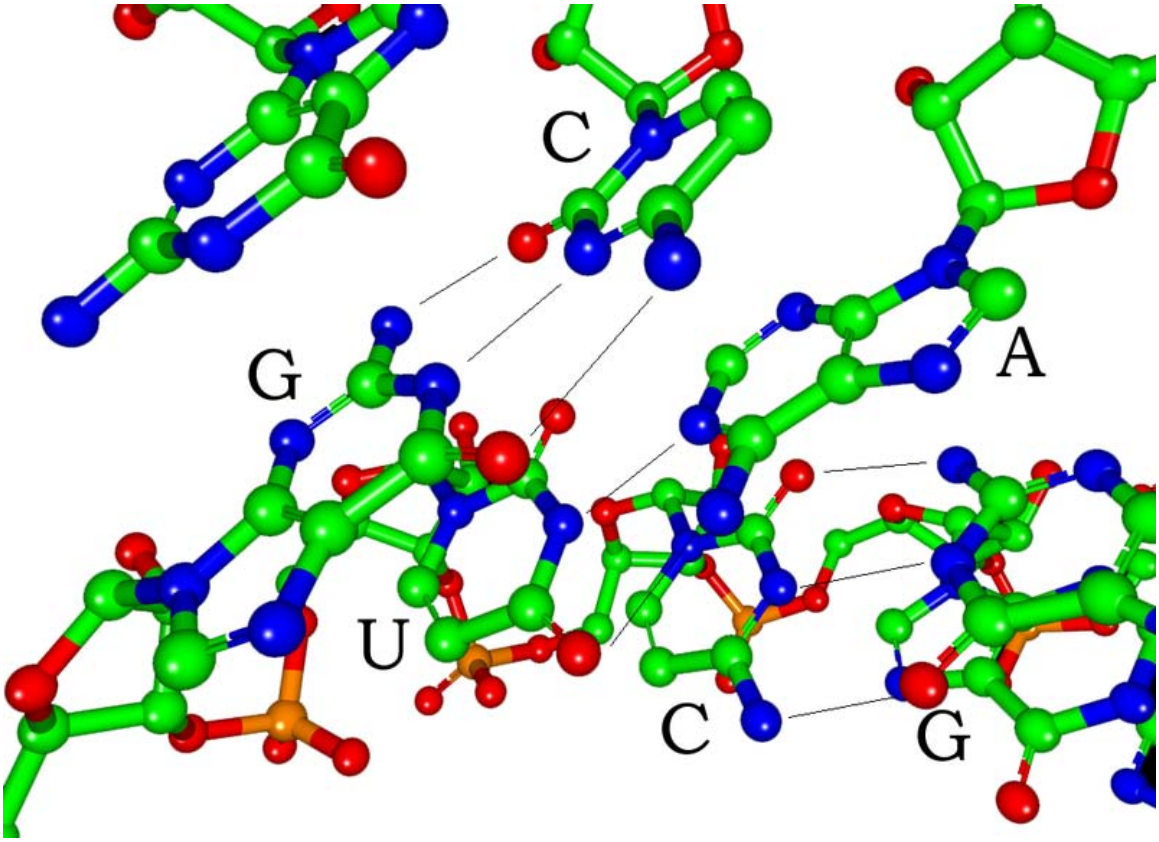
Three-dimensional representation of the 50S ribosomal subunit. RNA is in ochre, protein in blue. The active site is in the middle (red).

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

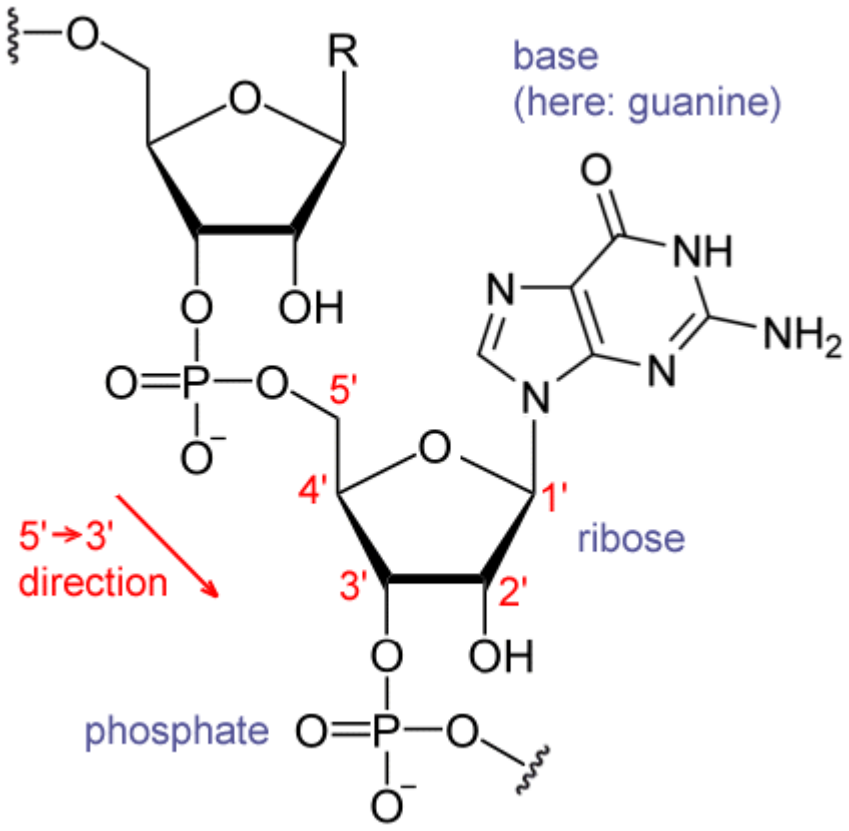
RNAs has revealed that they are highly structured. Unlike DNA, their structures do not consist of long double helices but rather collections of short helices packed together into structures akin to proteins. In this fashion, RNAs can achieve chemical catalysis, like enzymes. For instance, determination of the structure of the ribosome—an enzyme that catalyzes peptide bond formation—revealed that its active site is composed entirely of RNA.

Structure



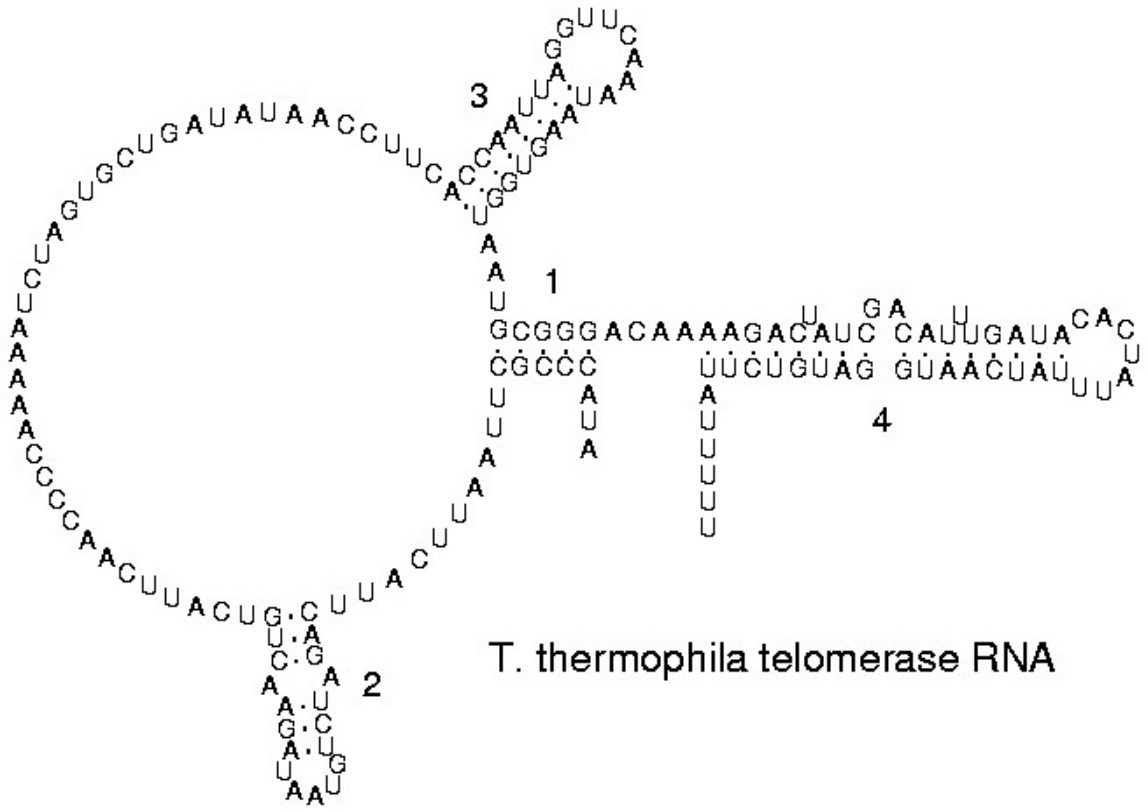
Watson-Crick base pairs in a siRNA (hydrogen atoms are not shown)

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

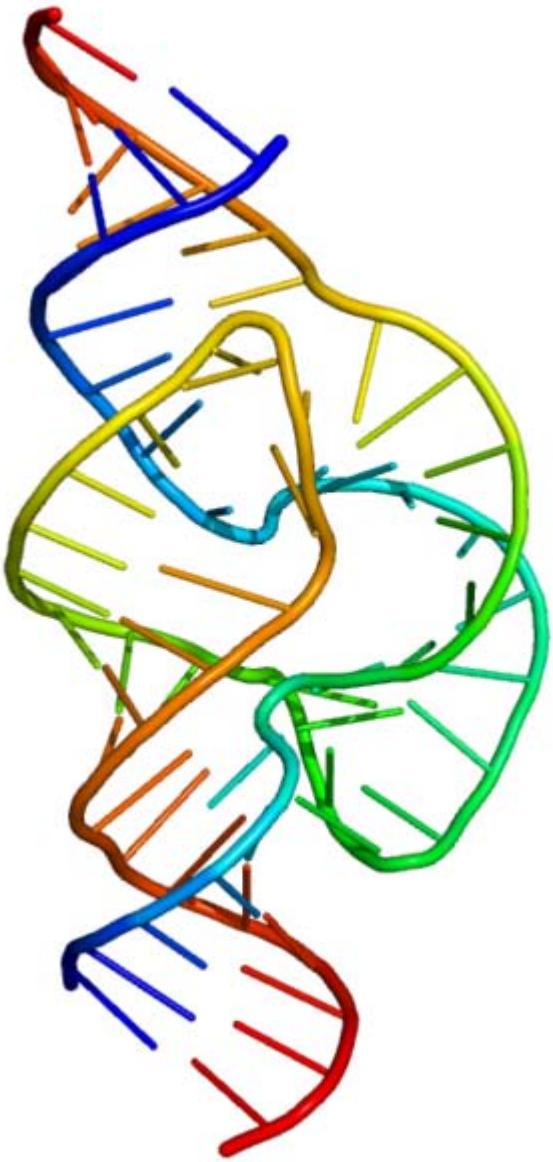
Synthesis of RNA is usually catalyzed by an enzyme—RNA polymerase—using DNA as a template, a process known as transcription. Initiation of transcription begins with the binding of the enzyme to a promoter sequence in the DNA (usually found "upstream" of a gene). The DNA double helix is unwound by the 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

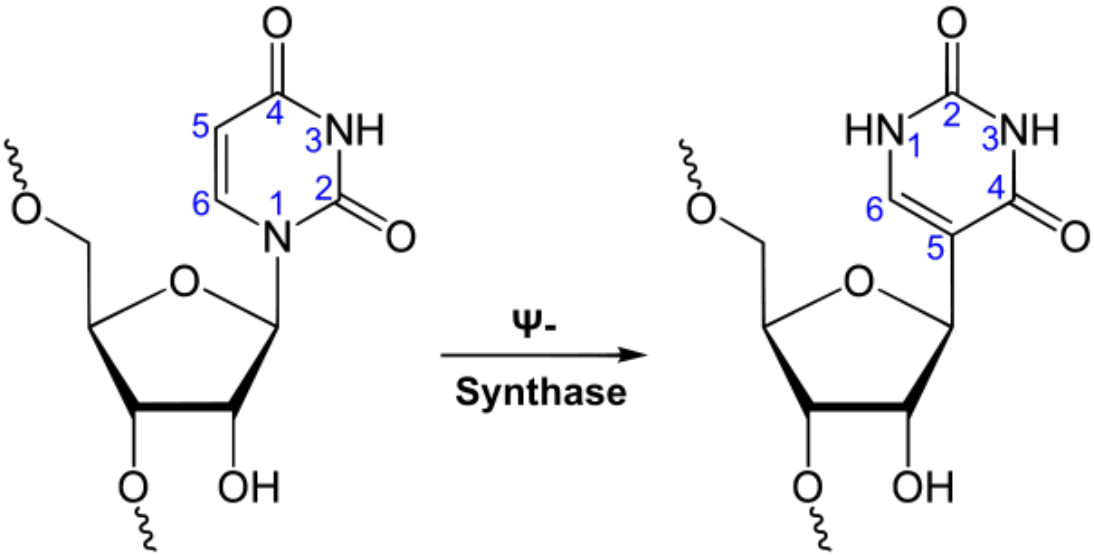
interfering RNAs (siRNA; 20-25 nt) are often produced by breakdown of viral RNA, there are also endogenous sources of siRNAs.

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.

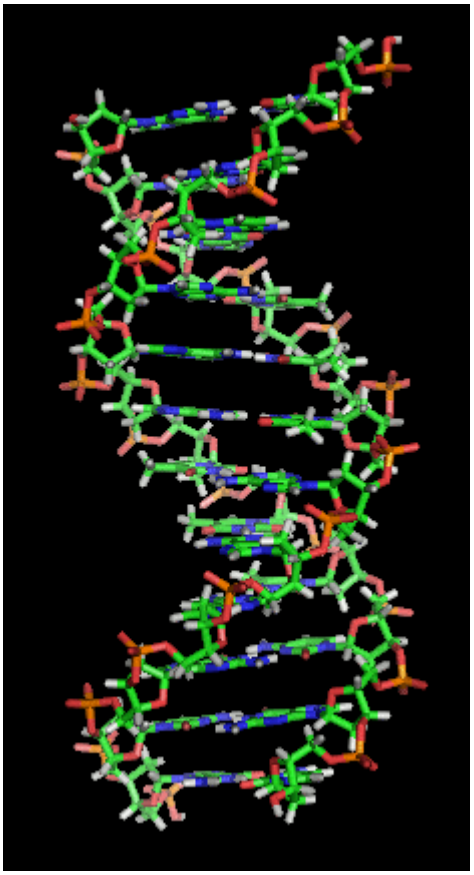
During the early 1970s, retroviruses and reverse transcriptase were discovered, showing for the first time that enzymes could copy RNA into DNA (the opposite of the usual route for transmission of genetic information). For this work, David Baltimore, Renato Dulbecco and Howard Temin were awarded a Nobel 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 6

Nucleic Acid Double Helix



Two complementary regions of nucleic acid molecules will bind and form a double helical structure held together by base pairs.

In molecular biology, the term **double helix** refers to the structure formed by double-stranded molecules of nucleic acids such as DNA and RNA. The double helical structure of a nucleic acid complex arises as a consequence of its secondary structure, and is a fundamental component in determining its tertiary structure.

The DNA double helix is a spiral polymer of nucleic acids, held together by nucleotides which base pair together. In B-DNA, the most common double helical structure, the double helix is right-handed with about 10–10.5 nucleotides per turn. The double helix structure of DNA contains a **major groove** and **minor groove**, the major groove being wider than the minor groove. Given the difference in widths of the major groove and minor groove, many proteins which bind to DNA do so through the wider major groove.

History

The double-helix model of DNA structure was first published in the journal Nature by James D. Watson and Francis Crick in 1953, based upon the crucial X-ray diffraction image of DNA labeled as "Photo 51", from Rosalind Franklin in 1952, followed by her more clarified DNA image with Raymond Gosling, Maurice Wilkins, Alexander Stokes, and Herbert Wilson, as well as base-pairing chemical and biochemical information by Erwin Chargaff. The previous model was triple-stranded DNA.

Crick, Wilkins, and Watson each received one third of the 1962 Nobel Prize in Physiology or Medicine for their contributions to the discovery. (Franklin, whose breakthrough X-ray diffraction data was used to formulate the DNA structure, died in 1958, and thus was ineligible to be nominated for a Nobel Prize.)

Nucleic acid hybridization

Hybridization is the process of complementary base pairs binding to form a double helix. Melting is the process by which the interactions between the strands of the double helix are broken, separating the two nucleic acid strands. These bonds are weak, easily separated by gentle heating, enzymes, or physical force. Melting occurs preferentially at certain points in the nucleic acid. **T** and **A** rich sequences are more easily melted than **C** and **G** rich regions. Particular base steps are also susceptible to DNA melting, particularly **T A** and **T G** base steps. These mechanical features are reflected by the use of sequences such as **TATAA** at the start of many genes to assist RNA polymerase in melting the DNA for transcription.

Strand separation by gentle heating, as used in PCR, is simple providing the molecules have fewer than about 10,000 base pairs (10 kilobase pairs, or 10 kbp). The intertwining of the DNA strands makes long segments difficult to separate. The cell avoids this problem by allowing its DNA-melting enzymes (helicases) to work concurrently with topoisomerases, which can chemically cleave the phosphate backbone of one of the strands so that it can swivel around the other. Helicases unwind the strands to facilitate the advance of sequence-reading enzymes such as DNA polymerase.

Base pair geometry

The geometry of a base, or base pair step can be characterized by 6 coordinates: Shift, slide, rise, tilt, roll, and twist. These values precisely define the location and orientation in space of every base or base pair in a nucleic acid molecule relative to its predecessor

along the axis of the helix. Together, they characterize the helical structure of the molecule. In regions of DNA or RNA where the "normal" structure is disrupted, the change in these values can be used to describe such disruption.

For each base pair, considered relative to its predecessor, there are the following base pair geometries to consider:

- **Shear**
- **Stretch**
- **Stagger**
- **Buckle**
- **Propeller twist**: rotation of one base with respect to the other in the same base pair.
- **Opening**
- **Shift**: displacement along an axis in the base-pair plane perpendicular to the first, directed from the minor to the major groove.
- **Slide**: displacement along an axis in the plane of the base pair directed from one strand to the other.
- **Rise**: displacement along the helix axis.
- **Tilt**: rotation around this axis.
- **vRoll**: rotation around this axis.
- **Twist**: rotation around the helix axis.
- **vx-displacement**
- **y-displacement**
- **inclination**
- **tip**
- **pitch**: the number of base pairs per complete turn of the helix.

Rise and twist determine the handedness and pitch of the helix. The other coordinates, by contrast, can be zero. Slide and shift are typically small in B-DNA, but are substantial in A- and Z-DNA. Roll and tilt make successive base pairs less parallel, and are typically small. A diagram of these coordinates can be found in 3DNA website.

Note that "tilt" has often been used differently in the scientific literature, referring to the deviation of the first, inter-strand base-pair axis from perpendicularity to the helix axis. This corresponds to slide between a succession of base pairs, and in helix-based coordinates is properly termed "inclination".

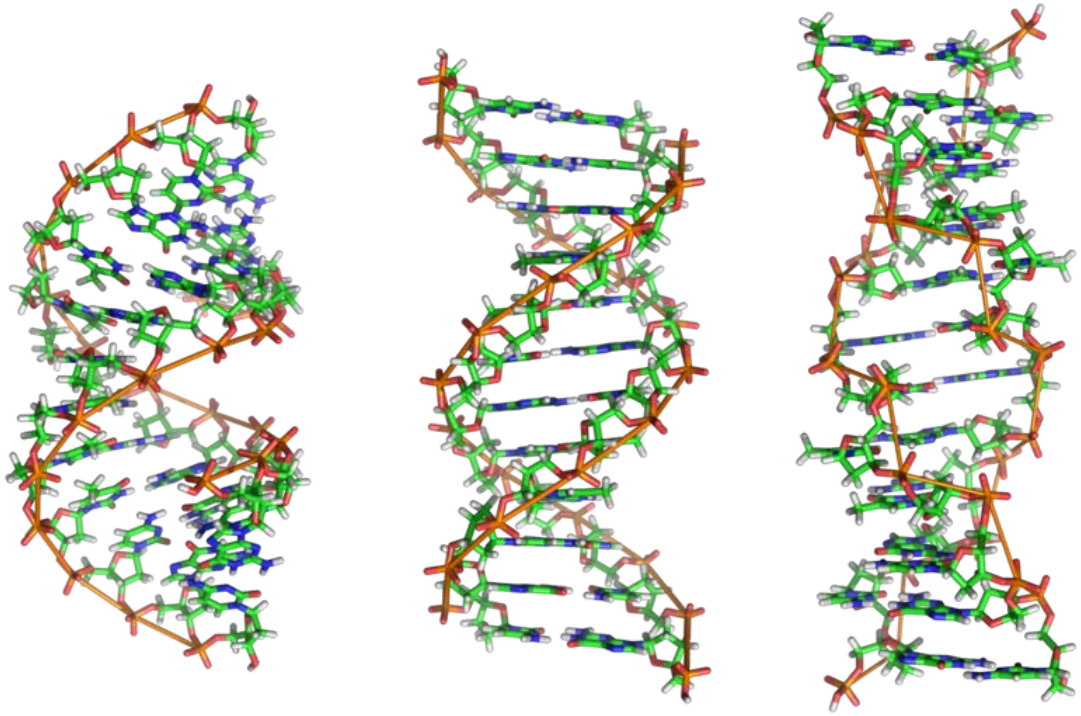
DNA helix geometries

At least three DNA conformations are believed to be found in nature, A-DNA, B-DNA, and Z-DNA. The "B" form described by James D. Watson and Francis Crick is believed to predominate in cells. It is 23.7 Å wide and extends 34 Å per 10 bp of sequence. The double helix makes one complete turn about its axis every 10.4-10.5 base pairs in solution. This frequency of twist (known as the helical *pitch*) depends largely on stacking forces that each base exerts on its neighbours in the chain.

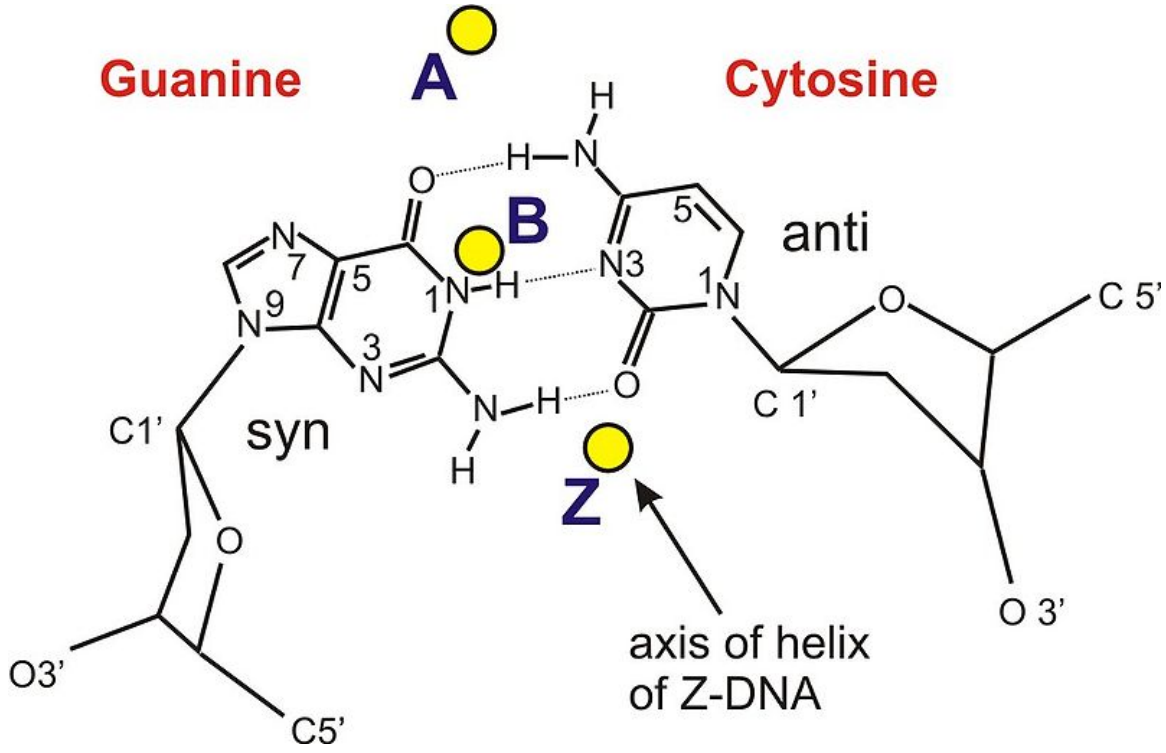
Other conformations are possible; A-DNA, B-DNA, C-DNA, E-DNA, L-DNA (the enantiomeric form of D-DNA), P-DNA, S-DNA, Z-DNA, etc. have been described so far. In fact, only the letters F, Q, U, V, and Y are now available to describe any new DNA structure that may appear in the future. However, most of these forms have been created synthetically and have not been observed in naturally occurring biological systems. Also note the triple-stranded DNA possibility.

A- and Z-DNA

A-DNA and Z-DNA differ significantly in their geometry and dimensions to B-DNA, although still form helical structures. The A form appears likely to occur only in dehydrated samples of DNA, such as those used in crystallographic experiments, and possibly in hybrid pairings of DNA and RNA strands. Segments of DNA that cells have methylated for regulatory purposes may adopt the Z geometry, in which the strands turn about the helical axis the opposite way to A-DNA and B-DNA. There is also evidence of protein-DNA complexes forming Z-DNA structures.



The structures of A-, B-, and Z-DNA.



The helix axis of A-, B-, and Z-DNA.

Structural features of the three major forms of DNA

Geometry attribute	A-DNA	B-DNA	Z-DNA
Helix sense	right-handed	right-handed	left-handed
Repeating unit	1 bp	1 bp	2 bp
Rotation/bp	32.7°	35.9°	60°/2
bp/turn	11	10.5	12
Inclination of bp to axis	+19°	−1.2°	−9°
Rise/bp along axis	2.3 Å (0.23 nm)	3.32 Å (0.332 nm)	3.8 Å (0.38 nm)
Pitch/turn of helix	28.2 Å (2.82 nm)	33.2 Å (3.32 nm)	45.6 Å (4.56 nm)
Mean propeller twist	+18°	+16°	0°
Glycosyl angle	anti	anti	C: anti, G: syn
Sugar pucker	C3'-endo	C2'-endo	C: C2'-endo, G: C2'-exo
Diameter	23 Å (2.3 nm)	20 Å (2.0 nm)	18 Å (1.8 nm)

Supercoiled DNA

The B form of the DNA helix twists 360° per 10.4-10.5 bp in the absence of torsional strain. But many molecular biological processes can induce torsional strain. A DNA

segment with excess or insufficient helical twisting is referred to, respectively, as positively or negatively "supercoiled". DNA *in vivo* is typically negatively supercoiled, which facilitates the unwinding (melting) of the double-helix required for RNA transcription.

Non-double-helical forms

Other non-double helical forms of DNA have been described, for example side-by-side (SBS) and triple helical configurations. Single stranded DNA may exist *in statu nascendi* or as thermally induced despiralized DNA.

Side-by-side models of DNA were proposed early in the history of molecular biology, but these were dropped in favor of the double-helical model due to X-ray crystallography of DNA duplexes as well as the nucleosome core particle, as well as the discovery of topoisomerases.

DNA bending

DNA is a relatively rigid polymer, typically modelled as a worm-like chain. It has three significant degrees of freedom; bending, twisting and compression, each of which cause particular limitations on what is possible with DNA within a cell. Twisting/torsional stiffness is important for the circularisation of DNA and the orientation of DNA bound proteins relative to each other and bending/axial stiffness is important for DNA wrapping and circularisation and protein interactions. Compression/extension is relatively unimportant in the absence of high tension.

Example sequences and their persistence lengths (B DNA)

Sequence	Persistence Length /base pairs
Random	154±10
$(\mathbf{CA})_{\text{repeat}}$	133±10
$(\mathbf{CAG})_{\text{repeat}}$	124±10
$(\mathbf{TATA})_{\text{repeat}}$	137±10

Persistence length/Axial stiffness

DNA in solution does not take a rigid structure but is continually changing conformation due to thermal vibration and collisions with water molecules, which makes classical measures of rigidity impossible. Hence, the bending stiffness of DNA is measured by the persistence length, defined as:

> "The length of DNA over which the time-averaged orientation of the polymer becomes uncorrelated by a factor of *e*".

This value may be directly measured using an atomic force microscope to directly image DNA molecules of various lengths. In an aqueous solution, the average persistence length is 46-50 nm or 140-150 base pairs (the diameter of DNA is 2 nm), although can vary significantly. This makes DNA a moderately stiff molecule.

The persistence length of a section of DNA is somewhat dependent on its sequence, and this can cause significant variation. The variation is largely due to base stacking energies and the residues which extend into the minor and major grooves.

Models for DNA bending

Stacking stability of base steps (B DNA)

Step	Stacking ΔG /kcal mol^{-1}
T A	-0.19
T G or **C A**	-0.55
C G	-0.91
A G or **C T**	-1.06
A A or **T T**	-1.11
A T	-1.34
G A or **T C**	-1.43
C C or **G G**	-1.44
A C or **G T**	-1.81
G C	-2.17

The entropic flexibility of DNA is remarkably consistent with standard polymer physics models such as the *Kratky-Porod* worm-like chain model. Consistent with the worm-like chain model is the observation that bending DNA is also described by Hooke's law at very small (sub-piconewton) forces. However for DNA segments less than the persistence length, the bending force is approximately constant and behaviour deviates from the worm-like chain predictions.

This effect results in unusual ease in circularising small DNA molecules and a higher probability of finding highly bent sections of DNA.

Bending preference

DNA molecules often have a preferred direction to bend, ie. anisotropic bending. This is, again, due to the properties of the bases which make up the DNA sequence - a random sequence will have no preferred bend direction, i.e. isotropic bending.

Preferred DNA bend direction is determined by the stability of stacking each base on top of the next. If unstable base stacking steps are always found on one side of the DNA helix then the DNA will preferentially bend away from that direction. As bend angle increases

then steric hindrances and ability to roll the residues relative to each other also play a role, especially in the minor groove. **A** and **T** residues will be preferentially be found in the minor grooves on the inside of bends. This effect is particularly seen in DNA-protein binding where tight DNA bending is induced, such as in nucleosome particles.

DNA molecules with exceptional bending preference can become intrinsically bent. This was first observed in trypanosomatid kinetoplast DNA. Typical sequences which cause this contain stretches of 4-6 **T** and **A** residues separated by **G** and **C** rich sections which keep the A and T residues in phase with the minor groove on one side of the molecule. For example:

```
 |                 |                  |                  |
|                 |
G A T T C C C A A A A A T G T C A A A A A A T A G G C A A A A A A T G C
C A A A A A A T C C C A A A C
```

The intrinsically bent structure is induced by the 'propeller twist' of base pairs relative to each other allowing unusual bifurcated Hydrogen-bonds between base steps. At higher temperatures this structure, and so the intrinsic bend, is lost.

All DNA which bends anisotropically has, on average, a longer persistence length and greater axial stiffness. This increased rigidity is required to prevent random bending which would make the molecule act isotropically.

DNA circularization

DNA circularization depends on both the axial (bending) stiffness and torsional (rotational) stiffness of the molecule. For a DNA molecule to successfully circularize it must be long enough to easily bend into the full circle and must have the correct number of bases so the ends are in the correct rotation to allow bonding to occur. The optimum length for circularization of DNA is around 400 base pairs (136 nm), with an integral number of turns of the DNA helix, i.e. multiples of 10.4 base pairs. Having a non integral number of turns presents a significant energy barrier for circularization, for example a 10.4 x 30 = 312 base pair molecule will circularize hundreds of times faster than 10.4 x 30.5 ≈ 317 base pair molecule.

DNA stretching

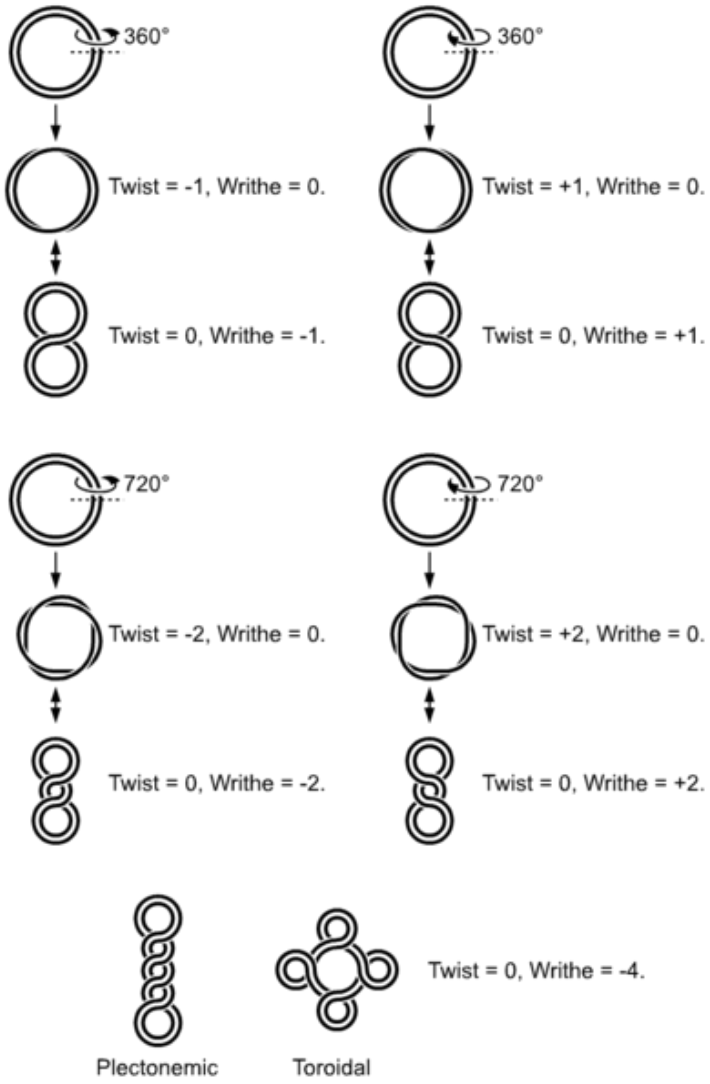
Longer stretches of DNA are entropically elastic under tension. When DNA is in solution, it undergoes continuous structural variations due to the energy available in the thermal bath of the solvent. This is due to the thermal vibration of the molecule combined with continual collisions with water molecules. For entropic reasons, more compact relaxed states are thermally accessible than stretched out states, and so DNA molecules are almost universally found in a tangled relaxed layouts. For this reason, a single molecule of DNA will stretch under a force, straightening it out. Using optical tweezers, the entropic stretching behavior of DNA has been studied and analyzed from a polymer

physics perspective, and it has been found that DNA behaves largely like the *Kratky-Porod* worm-like chain model under physiologically accessible energy scales.

Under sufficient tension and positive torque, DNA is thought to undergo a phase transition with the bases splaying outwards and the phosphates moving to the middle. This proposed structure for overstretched DNA has been called "P-form DNA," in honor of Linus Pauling who originally presented it as a possible structure of DNA

The mechanical properties DNA under compression have not been characterized due to experimental difficulties in preventing the polymer from bending under the compressive force.

DNA topology



Supercoiled structure of circular DNA molecules with low writhe. Note that the helical nature of the DNA duplex is omitted for clarity.

Within the cell most DNA is topologically restricted. DNA is typically found in closed loops (such as plasmids in prokaryotes) which are topologically closed, or as very long molecules whose diffusion coefficients produce effectively topologically closed domains. Linear sections of DNA are also commonly bound to proteins or physical structures (such as membranes) to form closed topological loops.

Francis Crick was one of the first to propose the importance of linking numbers when considering DNA supercoils. In a paper published in 1976, Crick outlined the problem as follows:

In considering supercoils formed by closed double-stranded molecules of DNA certain mathematical concepts, such as the linking number and the twist, are needed. The meaning of these for a closed ribbon is explained and also that of the writhing number of a closed curve. Some simple examples are given, some of which may be relevant to the structure of chromatin.

Analysis of DNA topology uses three values:

> L = linking number - the number of times one DNA strand wraps around the other. It is an integer for a closed loop and constant for a closed topological domain.
> T = twist - total number of turns in the double stranded DNA helix. This will normally tend to approach the number of turns that a topologically open double stranded DNA helix makes free in solution: number of bases/10.5, assuming there are no intercalating agents (e.g., chloroquine) or other elements modifying the stiffness of the DNA.
> W = writhe - number of turns of the double stranded DNA helix around the superhelical axis
> $L = T + W$ and $\Delta L = \Delta T + \Delta W$

Any change of T in a closed topological domain must be balanced by a change in W, and vice versa. This results in higher order structure of DNA. A circular DNA molecule with a writhe of 0 will be circular. If the twist of this molecule is subsequently increased or decreased by supercoiling then the writhe will be appropriately altered, making the molecule undergo plectonemic or toroidal superhelical coiling.

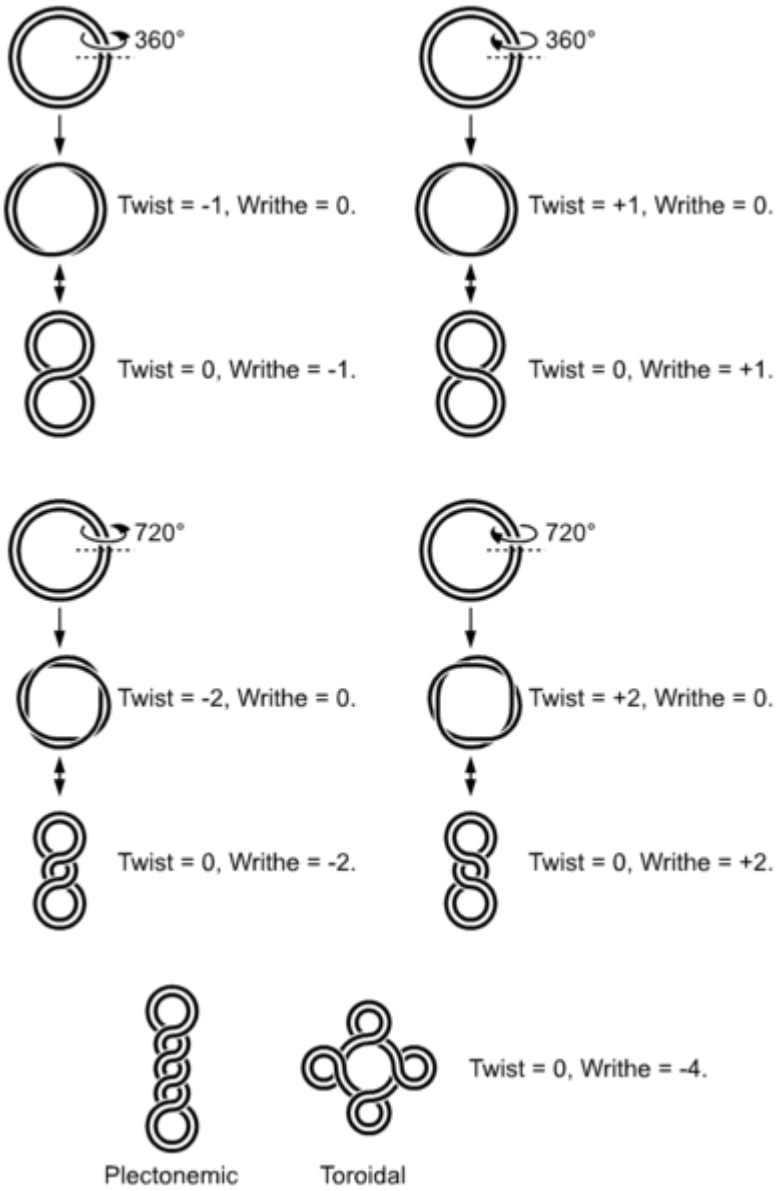
When the ends of a piece of double stranded helical DNA are joined so that it forms a circle the strands are topologically knotted. This means the single strands cannot be separated any process that does not involve breaking a strand (such as heating). The task of un-knotting topologically linked strands of DNA falls to enzymes known as topoisomerases. These enzymes are dedicated to un-knotting circular DNA by cleaving one or both strands so that another double or single stranded segment can pass through. This un-knotting is required for the replication of circular DNA and various types of recombination in linear DNA which have similar topological constraints.

The linking number paradox

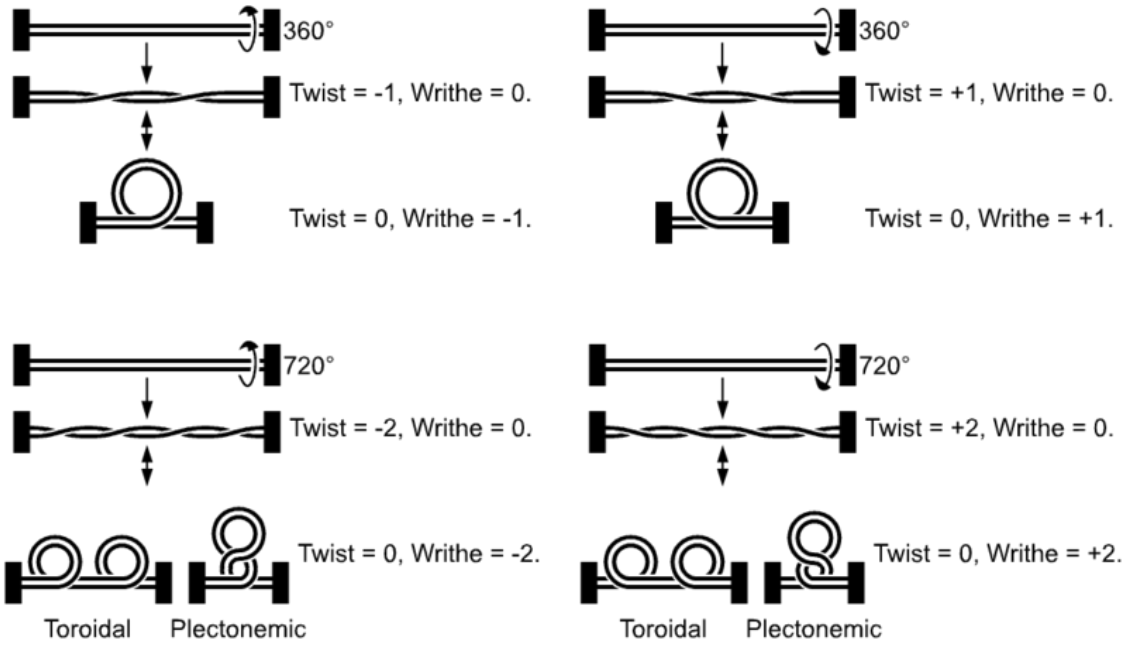
For many years, the origin of residual supercoiling in eukaryotic genomes remained unclear. This topological puzzle was referred to by some as the "linking number paradox". However, when experimentally determined structures of the nucleosome displayed an over-twisted left-handed wrap of DNA around the histone octamer, this "paradox" was solved.

Chapter 7

DNA Supercoil



Supercoiled structure of circular DNA molecules with low writhe. Note that the helical nature of the DNA duplex is omitted for clarity.



Supercoiled structure of linear DNA molecules with constrained ends. Note that the helical nature of the DNA duplex is omitted for clarity.

In a "relaxed" double-helical segment of **B-DNA**, the two strands twist around the helical axis once every 10.4-10.5 base pairs of sequence. Adding or subtracting twists, as some enzymes can do, imposes strain. If a DNA segment under twist strain were closed into a circle by joining its two ends and then allowed to move freely, the circular DNA would contort into a new shape, such as a simple figure-eight. Such a contortion is a **supercoil**.

The simple figure eight is the simplest supercoil, and is the shape a circular DNA assumes to accommodate one too many or one too few helical twists. The two lobes of the figure eight will appear rotated either clockwise or counterclockwise with respect to one another, depending on whether the helix is over or underwound. For each additional helical twist being accommodated, the lobes will show one more rotation about their axis.

The noun form "supercoil" is rarely used in the context of DNA topology. Instead, global contortions of a circular DNA, such as the rotation of the figure-eight lobes above, are referred to as *writhe*. The above example illustrates that twist and writhe are interconvertible. "**Supercoiling**" is an abstract mathematical property representing the sum of twist and writhe. The twist is the number of helical turns in the DNA and the writhe is the number of times the double helix crosses over on itself (these are the supercoils).

Extra helical twists are positive and lead to positive supercoiling, while subtractive twisting causes negative supercoiling. Many topoisomerase enzymes sense supercoiling

and either generate or dissipate it as they change DNA topology. DNA of most organisms is negatively supercoiled.

In part because chromosomes may be very large, segments in the middle may act as if their ends are anchored. As a result, they may be unable to distribute excess twist to the rest of the chromosome or to absorb twist to recover from underwinding--the segments may become *supercoiled*, in other words. In response to supercoiling, they will assume an amount of writhe, just as if their ends were joined.

Supercoiled DNA forms two structures; a plectoneme or a toroid, or a combination of both. A negatively supercoiled DNA molecule will produce either a one-start left-handed helix, the toroid, or a two-start right-handed helix with terminal loops, the plectoneme. Plectonemes are typically more common in nature, and this is the shape most bacterial plasmids will take. For larger molecules it is common for hybrid structures to form - a loop on a toroid can extend into a plectoneme. If all the loops on a toroid extend then it becomes a branch point in the plectonemic structure.

Occurrence of DNA supercoiling

DNA supercoiling is important for DNA packaging within all cells. Because the length of DNA can be thousands of times that of a cell, packaging this genetic material into the cell or nucleus (in eukaryotes) is a difficult feat. Supercoiling of DNA reduces the space and allows for a lot more DNA to be packaged. In prokaryotes, plectonemic supercoils are predominant, because of the circular chromosome and relatively small amount of genetic material. In eukaryotes, DNA supercoiling exists on many levels of both plectonemic and solenoidal supercoils, with the solenoidal supercoiling proving most effective in compacting the DNA. Solenoidal supercoiling is achieved with histones to form a 10nm fiber. This fiber is further coiled into a 30nm fiber, and further coiled upon itself numerous times more.

DNA packaging is greatly increased during nuclear division events such as mitosis or meiosis, where DNA must be compacted and segregated to daughter cells. Condensins and cohesins are *Structural Maintenance of Chromosome* proteins that aid in the condensation of sister chromatids and the linkage of the centromere in sister chromatids. These SMC proteins induce positive supercoils.

Supercoiling is also required for DNA/RNA synthesis. Because DNA must be unwound for DNA/RNA polymerase action, supercoils will result. The region ahead of the polymerase complex will be unwound; this stress is compensated with positive supercoils ahead of the complex. Behind the complex, DNA is rewound and there will be **compensatory** negative supercoils. It is important to note that topoisomerases such as DNA gyrase (Type II Topoisomerase) play a role in relieving some of the stress during DNA/RNA synthesis.

Modeling using mathematics

DNA supercoiling can be described numerically by changes in the 'linking number' Lk. The linking number is the most descriptive property of supercoiled DNA. Lk_o, the number of turns in the relaxed (B type) DNA plasmid/molecule, is determined by dividing the total base pairs of the molecule by the relaxed bp/turn which, depending on reference is 10.4-10.5.

$$Lk_o = bp / 10.4$$

Lk is merely the number of crosses a single strand makes across the other in a planar projection. The topology of the DNA is described by the equation below in which the linking number is equivalent to the sum of TW, which is the number of twists or turns of the double helix, and Wr which is the number of coils or 'writhes'. If there is a closed DNA molecule, the sum of TW and Wr, or the linking number, does not change. However, there may be complementary changes in TW and Wr without changing their sum.

$$Lk = Tw + Wr$$

The change in the linking number, ΔLk, is the actual number of turns in the plasmid/molecule, Lk, minus the number of turns in the relaxed plasmid/molecule Lk_o.

$$\Delta Lk = Lk - Lk_o$$

If the DNA is negatively supercoiled $\Delta Lk < 0$. The negative supercoiling implies that the DNA is underwound.

A standard expression independent of the molecule size is the "specific linking difference" or "superhelical density" denoted σ. σ represents the number of turns added or removed relative to the total number of turns in the relaxed molecule/plasmid, indicating the level of supercoiling.

$$\sigma = \Delta Lk / Lk_o$$

The Gibbs free energy associated with the coiling is given by the equation below

$$\Delta G / N = 10RT\sigma^2$$

Examples

Since the linking number L of supercoiled DNA is the number of times the two strands are intertwined (and both strands remain covalently intact), L cannot change. The reference state (or parameter) L_0 of a circular DNA duplex is its relaxed state. In this state, its writhe $W = 0$. Since $L = T + W$, in a relaxed state $T = L$. Thus, if we have a 400

bp relaxed circular DNA duplex, $L \sim 40$ (assuming ~10 bp per turn in B-DNA). Then $T \sim 40$.

- Positively supercoiling:

  T = 0, W = 0, then L = 0
  T = +3, W = 0, then L = +3
  T = +2, W = +1, then L = +3

- Negatively supercoiling:

  T = 0, W = 0, then L = 0
  T = -3, W = 0, then L = -3
  T = -2, W = -1, then L = -3

Negative supercoils favor local unwinding of the DNA, allowing processes such as transcription, DNA replication, and recombination. Negative supercoiling is also thought to favour the transition between B-DNA and Z-DNA, and moderate the interactions of DNA binding proteins involved in gene regulation.

Chapter 8

Nucleic Acid Structure Determination

Structure probing of nucleic acids is the process by which biochemical techniques are used to determine nucleic acid structure. This analysis can be used to define the patterns which can infer the molecular structure, experimental analysis of molecular structure and function, and further understanding on development of smaller molecules for further biological research. Structure probing analysis can be done through many different methods, which include chemical probing, hydroxyl radical probing, SHAPE, nucleotide analog interference mapping (NAIM), and in-line probing.

RNA sequencing

While methods for each type of probing differ in steps, the probing of secondary structure involves certain steps in order to determine the structure. The first step involved is submitting the structural RNA to the probe of interest and incubating over a certain amount of time to allow the reaction to occur. The RNA is then transcribed using reverse transcriptase PCR, where this results in different lengths of bands due to the modification of the RNA at specific sites, which causes the reverse transcriptase to fall off. These bands are then run on a gel with sequencing data determined from RNA sequencing.

Chemical probing

RNA chemical probing can involve many different chemicals which serve to modify specific bases at certain sites to show certain locations available for specific modification type.

DMS

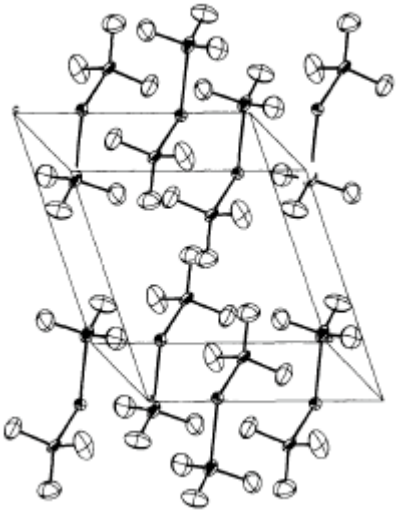


Figure 1. Structure of DMS showing the two methyl groups indicated as circles connected to sulfate.

Dimethyl sulfate, known as DMS, is a chemical that can be used to modify nucleic acids in order to determine secondary structure. DMS modifies certain bases through methylation. One set of methylation products that is used for RNA is the methylation of N1 adenosine and N3 of cytosine which prevents the natural hydrogen bonds to form between bases modified by DMS. This enables modification sites to be detected by RT-PCR, as modified sites cannot be basepaired, which results in the reverse transcriptase to fall off and different band sizes. DMS can only modify bases that are not basepaired, as single stranded nucleotides have the bases exposed so the chemical can modify. Detection of these modifications is done by examination of a gel and looking at bands present. One thing of note is that since modification prevents the addition of nucleotides at the modification site each of the bands generated is shifted by one base down on the gel. The PCR products indicate which adenine and cytosine nucleotides are double stranded, that is protected from DMS modification, or single stranded, that is accessible to DMS modification. This can be used to begin to develop a model of secondary structure for the RNA molecule. Structure probing by DMS allows for detection of secondary structure changes due to binding of RNA molecules along with detecting changes in tertiary structure, but it lacks the ability to determine the origin of these changes, as well as protection of the RNA backbone.

DMS modification can also be used for DNA, for example in footprinting DNA-protein interactions.

CMCT

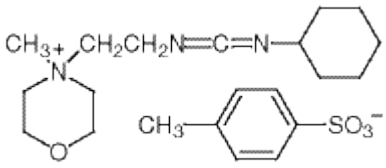


Figure 2. Structure of CMCT used in RNA structure probing

1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate known as CMCT is another chemical that is used in structure probing. CMCT like DMS serves to modify the exposed bases of specific nucleotides, which are uridine, and to a smaller extent guanine. CMCT reacts primarily with N3 of uridine and N1 of guanine modifying two sites responsible for hydrogen bonding on the bases. Modification using CMCT is analogous in detection to DMS as modification prevents basepairing at specific nucleotides which are then detected using rt-PCR and running of the PCR products on an agarose gel as shown in Figure 1. Structure probing of RNA by CMCT indicates the presence of uridine and guanine in single stranded regions by accessibility to modification or the presence of uridine and guanine in double stranded regions by protection from CMCT, which is the absence of a band. CMCT, like DMS can detect secondary and tertiary structure changes, but still has the same weaknesses as the method of modification is the same as DMS.

Kethoxal

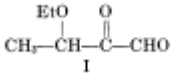


Figure 3. Structure of kethoxal that attacks guanine in structure probing.

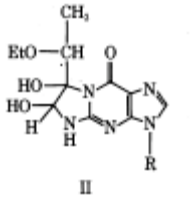


Figure 4. Binding of kethoxal to modified guanine preventing basepairing.

1,1-Dihydroxy-3-ethoxy-2-butanone, known as kethoxal, is used like DMS and CMCT, where treatment with kethoxal causes the modification of guanine, specifically altering the N1 and N2 by covalent interaction. The modification by kethoxal prevents single stranded guanine nucleotides to become modified, so that when reverse transcriptase reaches the modified guanine it falls off resulting in a band. After rt-PCR the products will be of different lengths where a band present indicates modification of a guanine base, and the absence of a band indicates that the base was not available for modification. Using this gel in combination with DMS and CMCT a model for structural RNAs can be

formed through comparison between the gels which indicate protected and accessible positions in all the bases, which can be used to form a preliminary model.

SHAPE

Selective 2′-**h**ydroxyl **a**cylation analyzed by **p**rimer **e**xtension, or **SHAPE**, takes advantage of reagents that preferentially modify the backbone of RNA in structurally flexible regions.

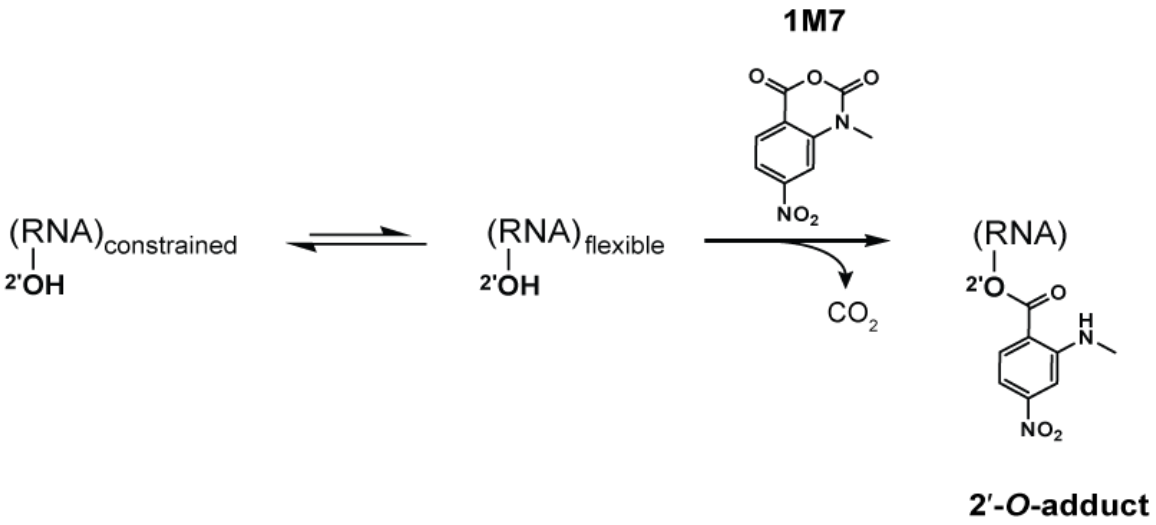


Figure 5. 1-methyl-7-nitroisatoic anhydride (1M7) undergoes hydrolysis to form adducts on the backbone of unpaired RNA nucleotides.

Reagents such as N-methylisotoic anhydride (NMIA) and 1-methyl-7-nitroisatoic anhydride (1M7) undergo hydrolysis to form adducts on the 2'-hydroxyl of the RNA backbone. Compared to the chemicals used in other RNA probing techniques, these reagents have the advantage of being largely unbiased to base identity, while remaining very sensitive to conformational dynamics. Nucleotides which are constrained (usually by base-pairing) show less adduct formation than nucleotides which are unpaired. Adduct formation is quantified for each nucleotide in a given RNA by extension of a complementary DNA primer with reverse transcriptase and comparison of the resulting fragments with those from an unmodified control. SHAPE therefore reports on RNA structure at the individual nucleotide level. This data can be used as input to generate highly accurate secondary structure models. SHAPE has been used to analyze diverse RNA structures, including that of an entire HIV-1 genome.

Hydroxyl radical probing

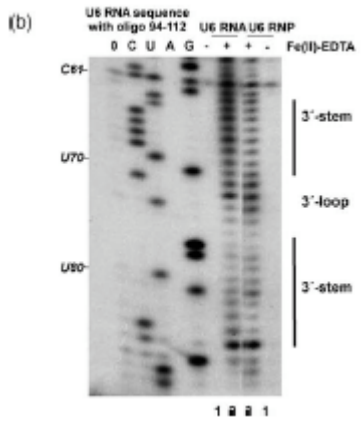


Figure 6. Hydroxyl radical probing gel showing bands at positions and dots indicating strength of protection.

Probing with hydroxyl radicals involves an additional step, as hydroxyl radicals are short lived in solution they need to be generated. This can be done using H_2O_2, ascorbic acid, and Fe(II)-EDTA complex which is attached to the backbone through EDTA. The Fe(II) along with ascorbic acid generates hydroxyl radicals which then can react with the nucleic acid molecules. Hydroxyl radical probing is often used in conjunction with chemical probing of nucleic acid molecules that are thought to associate with proteins, this is due to the modification done by hydroxyl radicals. Hydroxyl radicals attack the ribose/deoxyribose ring and this results in breaking of the phosphate backbone, which is independent of secondary structure, as all backbone is accessible, but is instead resultant of protein or tertiary structure protection. Probing with hydroxyl radicals shows the protection of structured nucleic acids by the proteins thought to be associated or folding on itself, where cleavage again results in a band formed through gel electrophoresis (after rt-PCR in the case of RNA) that is shorter than the full nucleic acid depending upon where it is cleaved. In this case since the last nucleotide is not modified the band length is indicative of the base that was cleaved. When examining the gel produced by running the gels on a band the areas of various strength of protection where areas of stronger protection for hydroxyl radicals can be said to have tighter association with a protein, or if no protein associates with the nucleic acid it can be caused by the tertiary fold.

In-line probing

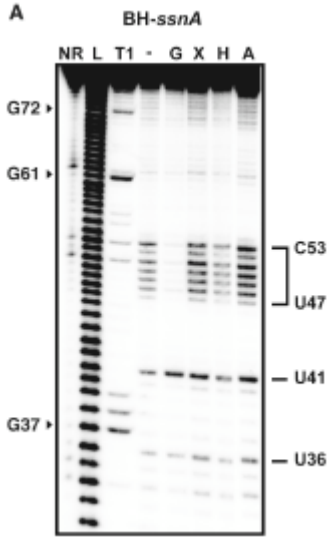


Figure 7. In-line probing assay of guanine riboswitches showing change in flexibility in response to various nucleotide ligands

In-line probing does not involve treatment with any type of chemicals or reagents to modify exposed RNA structures. This type of probing assay uses the structure dependent cleavage of RNA, as areas that are single stranded are more flexible and over time will degrade, as RNA structure is not always stable. The process of in-line probing is often used to determine changes in structure due to ligand binding as this can result in different cleavage patterns. The process of in-line probing involves incubation of structural or functional RNAs over a long period of time, can be several days, but varies in each experiment, and then running the incubated products on a gel to visualize the bands. This experiment is often done using two different conditions: 1) with ligand and 2) in the absence of ligand. Cleavage results in shorter band lengths and is indicative of areas that are not basepaired, as those that are tend to be less sensitive to spontaneous cleavage. In line probing is a functional assay in ligand binding changes of RNA because it can show directly the change in flexibility and binding of regions of DNA in response to a ligand, as well as compare that response to analogous ligands. An in-line probing assay is then commonly used in dynamic studies, specifically when examining riboswitches

Nucleotide analog interference mapping

Nucleotide analog interference mapping (NAIM) is the process of using nucleotide analogs, molecules that are similar in some ways to nucleotides but lack function, to determine the importance of a functional group at each location of an RNA molecule. The process of NAIM is to insert a single nucleotide analog into a unique site. This can be done by transcribing a short RNA using T7 RNA polymerase, then synthesizing a short oligonucleotide containing the analog in a specific position, then ligating them together on the DNA template using a ligase. The nucleotide analogs are tagged with a phosphorothioate, the active members of the RNA population are then distinguished from

the inactive members, the inactive members then have the phophorothioate tag removed and the analog sites are identified using gel electrophoresis and autoradiography. This indicates a functionally important nucleotides, as cleavage of the phosphorothioate by iodine results in an RNA that is cleaved at the site of the nucleotide analog insert. By running these truncated RNA molecules on a gel, the nucleotide of interest can be identified against a sequencing experiment Site directed incorporation results indicate positions of importance where when running on a gel, functional RNAs that have the analog incorporated at that position will have a band present, but if the analog results in non-functionality, when the functional RNA molecules are run on a gel there will be no band corresponding to that position on the gel. This process can be used to evaluate an entire area, where analogs are placed in site specific locations, differing by a single nucleotide, then when functional RNAs are isolated and run on a gel, all areas where bands are produced indicate non-essential nucleotides, but areas where bands are absent from the functional RNA indicate that inserting a nucleotide analog in that position caused the RNA molecule to become non-functional

Chapter 9

Nucleic Acid Structure Prediction

Nucleic acid structure prediction is a computational method to determine nucleic acid secondary and tertiary structure from its sequence. Secondary structure can be predicted from a single or from several nucleic acid sequences. Tertiary structure can be predicted from the sequence, or by comparative modeling (when the structure of a homologous sequence is known).

The problem of predicting nucleic acid secondary structure is dependent mainly on base pairing and base stacking interactions; many molecules have several possible three-dimensional structures, so predicting these structures remains out of reach unless obvious sequence and functional similarity to a known class of nucleic acid molecules, such as transfer RNA or microRNA, is observed. Many secondary structure prediction methods rely on variations of dynamic programming and therefore are unable to efficiently identify pseudoknots.

While the methods are similar, there are slight differences in the approaches to RNA and DNA structure prediction. *In vivo*, DNA structures are more likely to be duplexes with full complementarity between two strands, while RNA structures are more likely to fold into complex secondary and tertiary structures such as in the ribosome, spliceosome, or tRNA. This is partially because the extra oxygen in RNA increases the propensity for hydrogen bonding in the nucleic acid backbone. The energy parameters are also different for the two nucleic acids.

Single sequence structure prediction

A common problem for researchers working with RNA is to determine the three-dimensional structure of the molecule given just the nucleic acid sequence. However, in the case of RNA much of the final structure is determined by the secondary structure or intra-molecular base-pairing interactions of the molecule. This is shown by the high conservation of base-pairings across diverse species.

The most stable structure

Secondary structure of small RNA molecules is largely determined by strong, local interactions such as hydrogen bonds and base stacking. Summing the free energy for such interactions should provide an approximation for the stability of a given structure. To predict the folding free energy of a given secondary structure, an empirical nearest-neighbor model is used. In the nearest neighbor model the free energy change for each motif depends on the sequence of the motif and of its closest base-pairs. The model and parameters of minimal energy for Watson–Crick pairs, GU pairs and loop regions were derived from empirical calorimetric experiments, the most up-to-date parameters were published in 2004, although most software packages use the previous set assembled in 1999.

The simplest way to find the lowest free energy structure would be to generate all possible structures and calculate the free energy for it, but the number of possible structures for a sequence increases exponentially with the length of RNA (Number of secondary structures = $(1,8)^N$, N- number of nucleotides). For longer RNA molecules, the number of possible secondary structures is huge: a sequence of 100 nucleotides has more than 1025 possible secondary structures.

Dynamic programming algorithms

The first and the most popular method for finding the most stable structure is a dynamic programming algorithm. One of the first attempts to predict RNA secondary structure was made by Ruth Nussinov and co-workers who used the dynamic programming method for maximizing the number of base-pairs. However, there are several issues with this approach: most importantly, the solution is not unique. Nussinov et al. published an adaptation of their approach using a simple nearest-neighbour energy model in 1980. In 1981, Michael Zuker and Patrick Stiegler proposed using a slightly refined dynamic programming approach to modelling nearest neighbor energy interactions that directly incorporates stacking into the prediction.

Dynamic programming algorithms provide a means to implicitly check all variants of possible RNA secondary structures without explicitly generating the structures. First, the lowest conformational free energy is determined for each possible sequence fragment starting with the shortest fragments and then for longer fragments. For longer fragments, recursion on the optimal free energy changes determined for shorter sequences speeds the determination of the lowest folding free energy. Once the lowest free energy of the complete sequence is calculated, the exact structure of RNA molecule is determined.

Dynamic programming algorithms are commonly used to detect base pairing patterns that are "well-nested", that is, form hydrogen bonds only to bases that do not overlap one another in sequence position. Secondary structures that fall into this category include double helices, stem-loops, and variants of the "cloverleaf" pattern found in transfer RNA molecules. These methods rely on pre-calculated parameters which estimate the free energy associated with particular types of base-pairing interactions, including Watson-

Crick and Hoogsteen base pairs. Depending on the complexity of the method, single base pairs may be considered as well as short two- or three-base segments to incorporate the effects of base stacking. This method cannot identify pseudoknots, which are not well nested, without substantial algorithmic modifications that are extremely computationally expensive.

Suboptimal structures

The accuracy of RNA secondary structure prediction from a single sequence by free energy minimization is limited by several factors:

1. The free energy value's list in nearest neighbor model is incomplete
2. Not all known RNA folds in such a way as to conform with the thermodynamic minimum.
3. Some RNA sequences have more than one biologically active conformation (e. i. Riboswitches)

For this reason, the ability to predict structures which have similar low free energy would provide significant information. Such structures are termed suboptimal structures. MFOLD is one program that generates suboptimal structures.

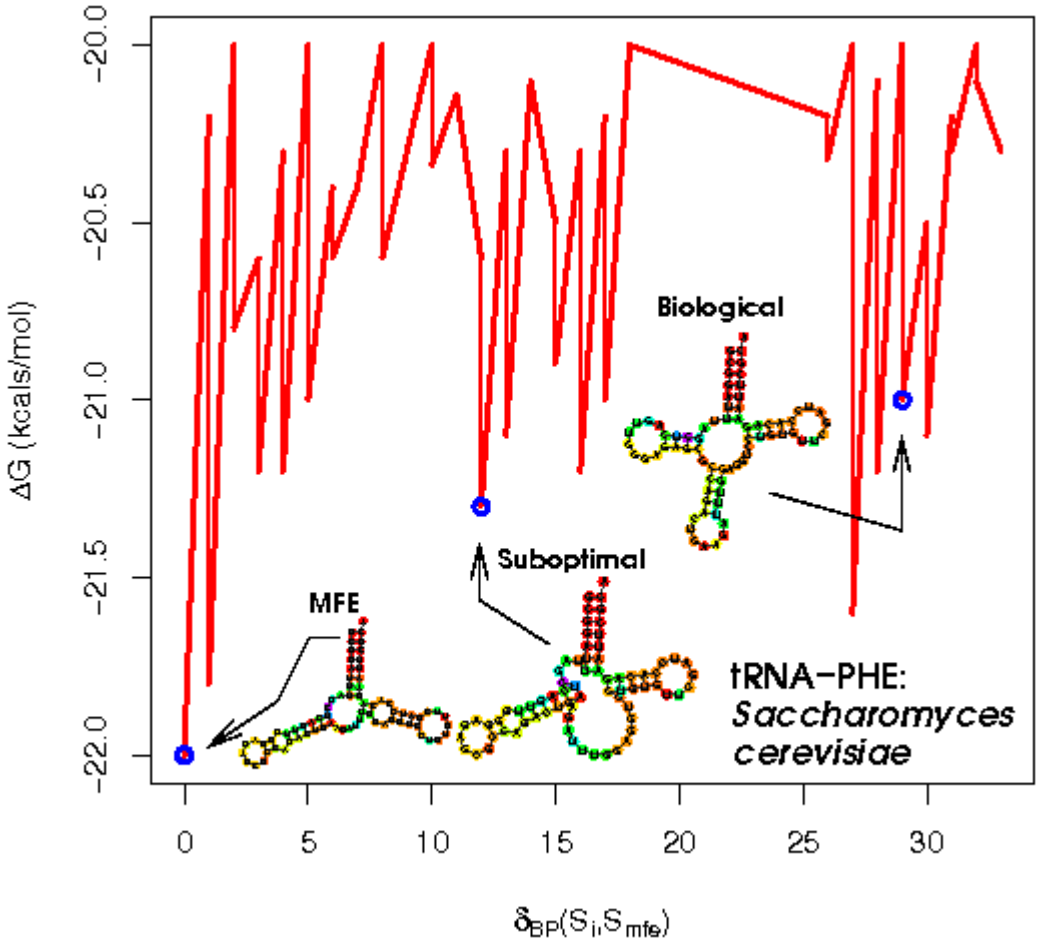
Predicting pseudoknots

One of the issues when predicting RNA secondary structure is that the standard free energy minimization and statistical sampling methods can not find pseudoknots. The major problem is that the usual dynamic programing algorithms, when predicting secondary structure, consider only the interactions between the closest nucleotides, while pseudoknotted structures are formed due to interactions between distant nucleotides. Rivas and Eddy published a dynamic programming algorithm for predicting pseudoknots. However, this dynamic programming algorithm is very slow. The standard dynamic programming algorithm for free energy minimization scales $O(N^3)$ in time (N is the number of nucleotides in the sequence), while the Rivas and Eddy algorithm scales $O(N^6)$ in time. This has prompted several researchers to implement versions of the algorithm that restrict classes of pseudoknots, resulting in performance gains. For example, pknotsRG tool includes only the class of simple recursive pseudoknots and scales O(N4) in time.

Other approaches for RNA secondary structure prediction

Another approach for RNA secondary structure determination is to sample structures from the Boltzmann ensemble, as exemplified by the program SFOLD. The program generates a statistical sample of all possible RNA secondary structures. The algorithm samples secondary structures according to the Boltzmann distribution. The sampling method offers an appealing solution to the problem of uncertainties in folding.

Comparative secondary structure prediction



S. cerevisiae tRNA-PHE structure space: the energies and structures were calculated using RNAsubopt and the structure distances computed using RNAdistance.

Sequence covariation methods rely on the existence of a data set composed of multiple homologous RNA sequences with related but dissimilar sequences. These methods analyze the covariation of individual base sites in evolution; maintenance at two widely separated sites of a pair of base-pairing nucleotides indicates the presence of a structurally required hydrogen bond between those positions. The general problem of pseudoknot prediction has been shown to be NP-complete.

In general, the problem of alignment and consensus structure prediction are closely related. Three different approaches to the prediction of consensus structures can be distinguished:

1. Folding of alignment
2. Simultaneous sequence alignment and folding
3. Alignment of predicted structures

Align then fold

A practical heuristic approach is to use multiple sequence alignment tools to produce an alignment of several RNA sequences, to find consensus sequence and then fold it. The quality of the alignment determines the accuracy of the consensus structure model. Consensus sequences are folded using various approaches similarly as in individual structure prediction problem. The thermodynamic folding approach is exemplified by RNAalifold program. The different approaches are exemplified by Pfold and ILM programs. Pfold program implements a SCFGs. ILM (iterated loop matching) unlike the other algorithms for folding of alignments, can return pseudocnoted structures. It uses combination of thermodynamics and mutual information content scores.

Align and fold

Evolution frequently preserves functional RNA structure better than RNA sequence. Hence, a common biological problem is to infer a common structure for two or more highly diverged but homologous RNA sequences. In practice, sequence alignments become unsuitable and do not help to improve the accuracy of structure prediction, when sequence similarity of two sequences is less than 50%.

Structure-based alignment programs improves the performance of these alignments and most of them are variants of the Sankoff algorithm. Basically, Sankoff algorithm is a merger of sequence alignment and Nussinov (maximal-pairing) folding dynamic programming method. Sankoff algorithm itself is a theoretical exercise because it requires extreme computational resources (($O^{(n3m)}$ in time, and $O^{(n2m)}$ in space, where n is the sequence length and m is the number of sequences). Some notable attempts at implementing restricted versions of Sankoff's algorithm are Foldalign, Dynalign, PMmulti/PMcomp, Stemloc, and Murlet. In these implementations the maximal length of alignment or variants of possible consensus structures are restricted. For example, Foldalign focuses on local alignments and restricts the possible length of the sequences alignment.

Fold then align

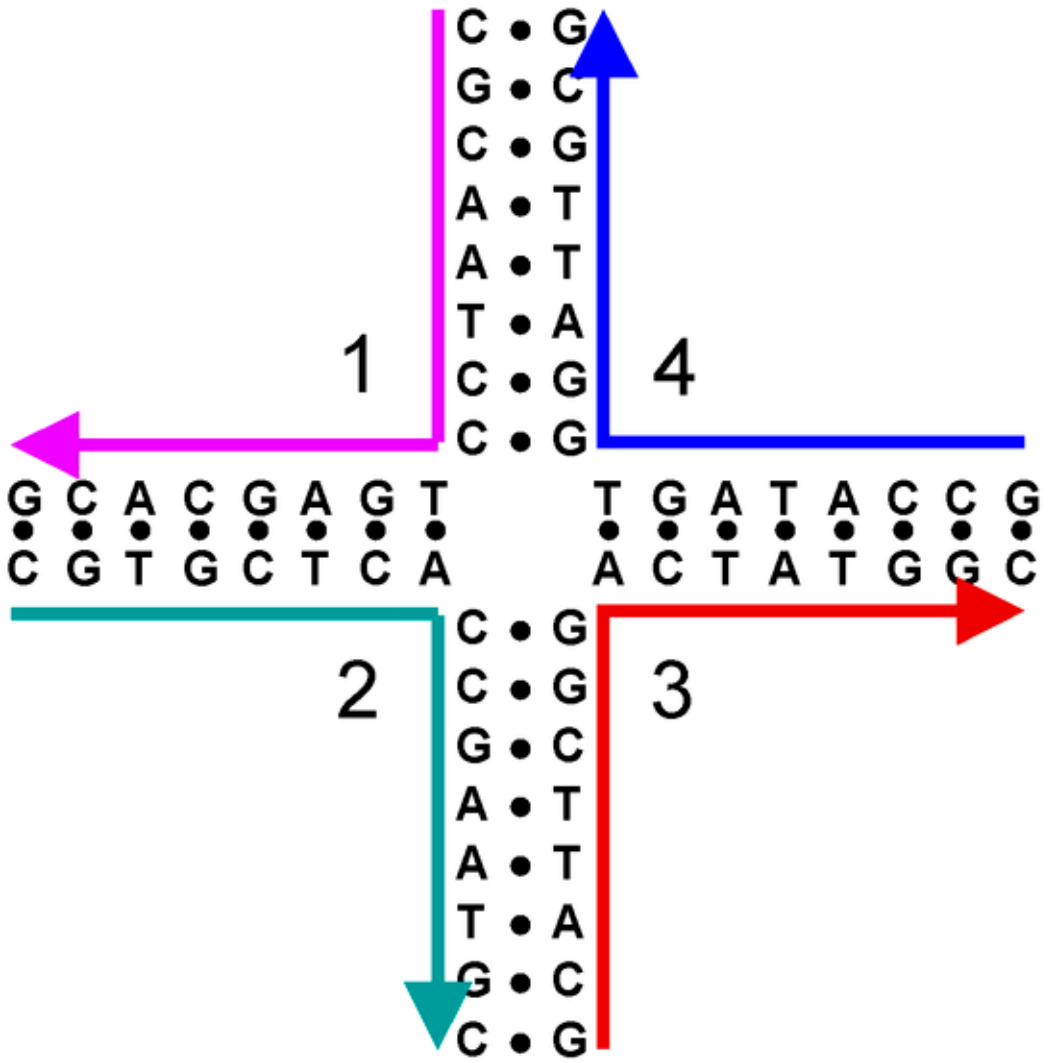
A less widely used approach is to fold the sequences using single sequence structure prediction methods and align the resulting structures using tree-based metrics. The fundamental weakness with this approach is that single sequence predictions are often inaccurate, thus all further analyses are affected.

Tertiary structure prediction

Once secondary structure of RNA is known, the next challenge is to predict tertiary structure. The biggest problem is to determine the structure of regions between double stranded helical regions. Also RNA molecules often contain posttranscriptionally modified nucleosides, which because of new possible non-canonical interactions, cause a lot of troubles for tertiary structure prediction.

Chapter 10

Nucleic Acid Design



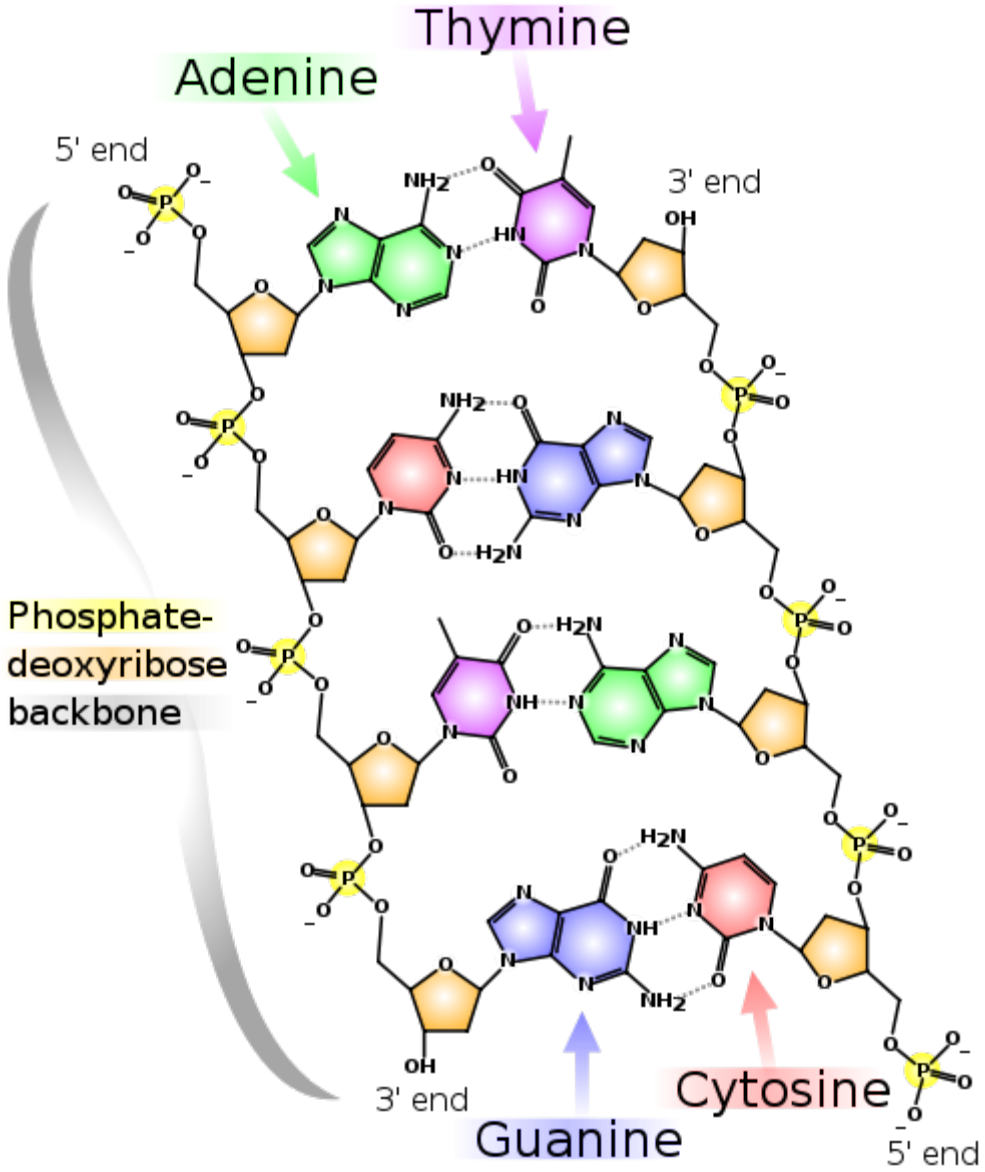
Nucleic acid design can be used to create nucleic acid complexes with complicated secondary structures such as this four-arm junction. Image from Mao, 2004.

Nucleic acid design is the process of generating a set of nucleic acid base sequences that will associate into a desired conformation. Nucleic acid design is central to the fields of DNA nanotechnology and DNA computing. It is necessary because there are many possible sequences of nucleic acid strands that will fold into a given secondary structure, but many of these sequences will have undesired additional interactions which must be avoided. In addition, there are many tertiary structure considerations which affect the choice of a secondary structure for a given design.

Nucleic acid design has similar goals to protein design: in both, the sequence of monomers is designed to favor the desired folded or associated structure and to disfavor alternate structures. However, nucleic acid design has the advantage of being a much computationally simpler problem, since the simplicity of Watson-Crick base pairing rules leads to simple heuristic methods which yield experimentally robust designs. Computational models for protein folding require tertiary structure information whereas nucleic acid design can operate largely on the level of secondary structure. However, nucleic acid structures are less versatile than proteins in their functionality.

Nucleic acid design can be considered the inverse of nucleic acid structure prediction. In structure prediction, the structure is determined from a known sequence, while in nucleic acid design, a sequence is generated which will form a desired structure.

Fundamental concepts



Chemical structure of DNA. Nucleic acid double helices will only form between two strands of complementary sequences, where the bases are matched into only A-T or G-C pairs.

The structure of nucleic acids consists of a sequence of nucleotides. There are four types of nucleotides distinguished by which of the four nucleobases they contain: in DNA these are adenine (A), cytosine (C), guanine (G), and thymine (T). Nucleic acids have the property that two molecules will bind to each other to form a double helix only if the two sequences are complementary, that is, they can form matching sequences of base pairs. Thus, in nucleic acids the sequence determines the pattern of binding and thus the overall structure.

Nucleic acid design is the process by which, given a desired target structure or functionality, sequences are generated for nucleic acid strands which will self-assemble into that target structure. Nucleic acid design encompasses all levels of nucleic acid structure:

- Primary structure—the raw sequence of nucleobases of each of the component nucleic acid strands;
- Secondary structure—the set of interactions between bases, i.e., which parts of which strands are bound to each other; and
- Tertiary structure—the locations of the atoms in three-dimensional space, taking into consideration geometrical and steric constraints.

One of the greatest concerns in nucleic acid design is ensuring that the target structure has the lowest energy (i.e. is the most thermodynamically favorable) whereas misformed structures have higher energy and are thus unfavored. These goals can be achieved through the use of a number of apporaches, including heuristic, thermodynamic, and geometrical ones. Almost all nucleic acid design tasks are aided by computers, and a number of software packages are available for many of these tasks.

Two considerations in nucleic acid design are that desired hybridizations should have melting temperatures in a narrow range, and any spurious interactions should have very low melting temperatures (i.e. they should be very weak). There is also a contrast between affinity-optimizing "positive design", seeks to minimize the energy of the desired structure in an absolute sense, and specificity-optimizing "negative design", which considers the energy of the target structure relative to those of undesired structures. Algorithms which implement both kinds of design tend to perform better than those that consider only one type.

Approaches

Heuristic methods

Heuristic methods use simple criteria which can be quickly evaluated to judge the suitability of different sequences for a given secondary structure. They have the advantage of being much less computationally expensive than the energy minimization algorithms needed for thermodynamic or geometrical modeling, and being easier to implement, but at the cost of being less rigorous than these models.

Sequence symmetry minimization is the oldest approach to nucleic acid design and was first used to design immobile versions of branched DNA structures. Sequence symmetry minimization divides the nucleic acid sequence into overlapping subsequences of a fixed length, called the criterion length. Each of the 4^N possible subsequences of length N is allowed to appear only once in the sequence. This ensures that no undesired hybridizations can occur which have a length greater than or equal to the criterion length.

A related heuristic approach is to consider the "mismatch distance", meaning the number of positions in a certain frame where the bases are not complementary. A greater mismatch distance lessens the chance that a strong spurious interaction can happen. This is related to the concept of Hamming distance in information theory. Another related but more involved approach is to use methods from coding theory to construct nucleic acid sequences with desired properties.

Thermodynamic models

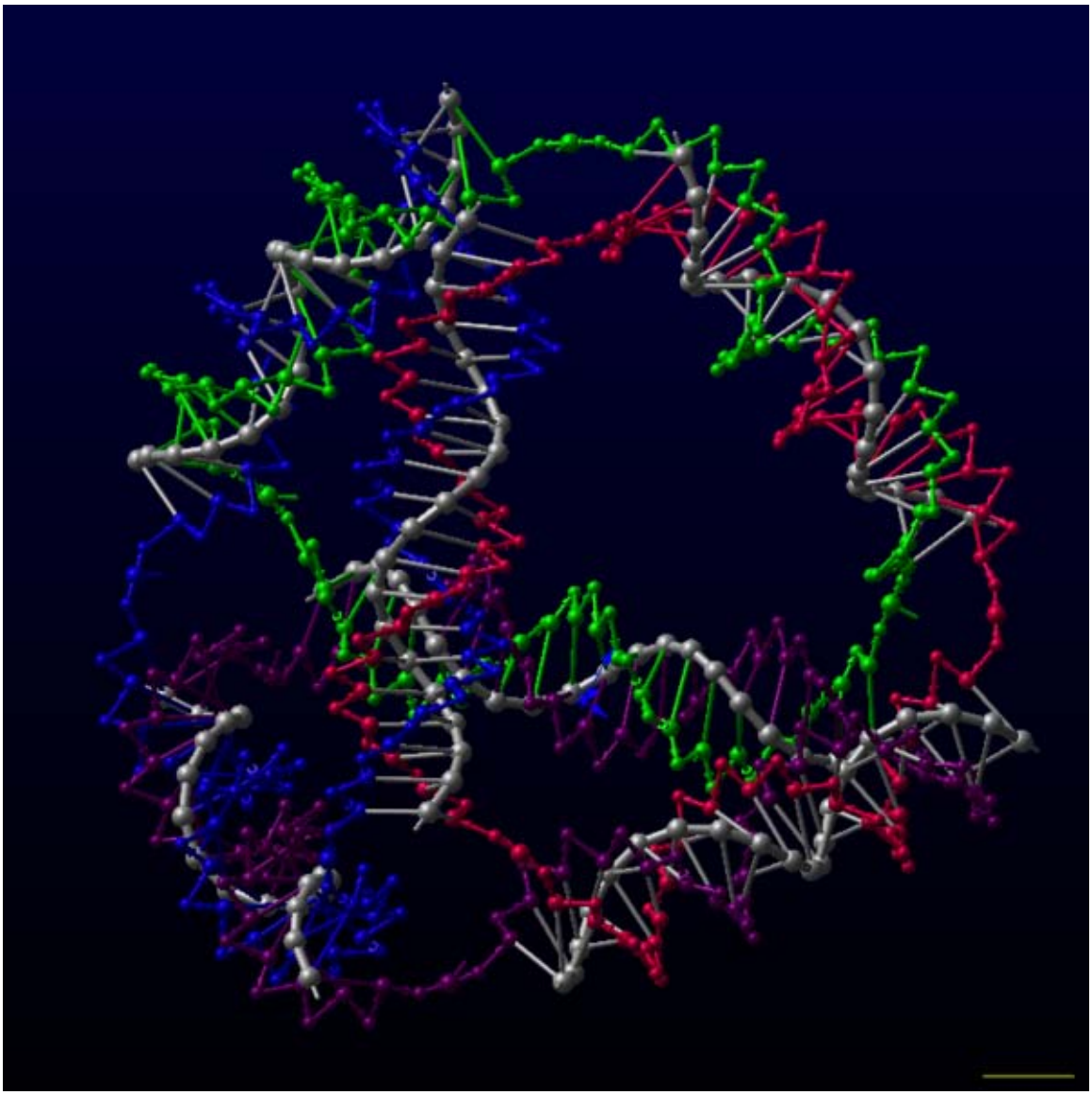
Information about the secondary structure of a nucleic acid complex along with its sequence can be used to predict the thermodyamic properties of the complex.

When thermodynamic models are used in nucleic acid design, there are usually two considerations: desired hybridizations should have melting temperatures in a narrow range, and any spurious interactions should have very low melting temperatures (i.e. they should be very weak). The free energy of a perfectly matched nucleic acid duplex can be predicted using a nearest neighbor model. This model considers only the interactions between a nucleotide and its nearest neighbors on the nucleic acid strand, by summing the free energy of each of the overlapping two-nucleotide subwords of the duplex. This is then corrected for self-complementary monomers and for GC-content. Once the free energy is known, the melting temperature of the duplex can be determined. GC-content alone can also be used to estimate the free energy and melting temperature of a nucleic acid duplex. This is less accurate but also much less computationally costly.

Software for thermodynamic modeling of nucleic acids includes Nupack, mfold, and Vienna.

A related approach, inverse secondary structure prediction, uses stochastic local search which improves a nucleic acid sequence by running a structure prediction algorithm and the modifying the sequence to eliminate unwanted features.

Geometrical models



A geometrical model of a DNA tetrahedron described in Goodman, 2005. Models of this type are useful for ensuring that tertiary structure constraints do not cause excessive strain to the molecule.

Geometrical models of nucleic acids are used to predict tertiary structure. This is important because designed nucleic acid complexes usually contain multiple junction points, which introduces geometric constraints to the system. These constraints stem from the basic structure of nucleic acids, mainly that the double helix formed by nucleic acid duplexes has a fixed helicity of about 10.4 base pairs per turn, and is relatively stiff. Because of these constraints, the nucleic acid complexes are sensitive to the relative orientation of the major and minor grooves at junction points. Geometrical modeling can detect strain stemming from misalignments in the structure, which can then be corrected by the designer.

Geometric models of nucleic acids for DNA nanotechnology generally use reduced representations of the nucleic acid, because simulating every atom would be very computationally expensive for such large systems. Models with three pseudo-atoms per base pair, representing the two backbone sugars and the helix axis, have been reported to have a sufficient level of detail to predict experimental results. However, models with five pseudo-atoms per base pair, explicitly including the backbone phosphates, are also used.

Software for geometrical modeling of nucleic acids includes GIDEON, Tiamat, Nanoengineer-1, and UNIQUIMER 3D. Geometrical concerns are especially of interest in the design of DNA origami, because the sequence is predetermined by the choice of scaffold strand. Software specifically for DNA origami design has been made, including caDNAno and SARSE.

Applications

Nucleic acid design is used in DNA nanotechnology to design strands which will self-assemble into a desired target structure. These include examples such as DNA machines, periodic two- and three-dimensional lattices, polyhedra, and DNA origami. It can also be used to create sets of nucleic acid strands which are "orthogonal", or non-interacting with each other, so as to minimize or eliminate spurious interactions. This is useful in DNA computing, as well as for molecular barcoding applications in chemical biology and biotechnology.

Chapter 11

Nucleic Acid Thermodynamics

Nucleic acid thermodynamics is the study of the thermodynamics of nucleic acid molecules, or how temperature affects nucleic acid structure. For multiple copies of DNA molecules, the melting temperature (T_m) is defined as the temperature at which half of the DNA strands are in the double-helical state and half are in the single stranded states. The melting temperature depends on both the length of the molecule, and the specific nucleotide sequence composition of that molecule.

Concepts

Hybridization

Hybridization is the process of establishing a non-covalent, sequence-specific interaction between two or more complementary strands of nucleic acids into a single hybrid, which in the case of two strands is referred to as a duplex. Oligonucleotides, DNA, or RNA will bind to their complement under normal conditions, so two perfectly complementary strands will bind to each other readily. In order to reduce the diversity and obtain the most energetically preferred hybrids, a technique called annealing is used in laboratory practice. However, due to the different molecular geometries of the nucleotides, a single inconsistency between the two strands will make binding between them less energetically favorable. Measuring the effects of base incompatibility by quantifying the rate at which two strands anneal can provide information as to the similarity in base sequence between the two strands being annealed. The hybrids may be dissociated by thermal denaturation, also referred to as melting. Here, the solution of hybrids is heated to break the hydrogen bonds between nucleic bases, after which the two strands separate. In the absence of external negative factors, the processes of hybridization and melting may be repeated in succession indefinitely, which lays the ground for polymerase chain reaction. Most commonly, the pairs of nucleic bases A=T and G≡C are formed, of which the latter is more stable.

Denaturation

DNA denaturation, also called **DNA melting**, is the process by which double-stranded deoxyribonucleic acid unwinds and separates into single-stranded strands through the breaking of hydrogen bonding between the bases. Both terms are used to refer to the process as it occurs when a mixture is heated, although "denaturation" can also refer to the separation of DNA strands induced by chemicals like urea .

The process of DNA denaturation can be used to analyze some aspects of DNA. Because cytosine / guanine base-pairing is generally stronger than adenosine / thymine base-pairing, the amount of cytosine and guanine in a genome (called the "GC content") can be estimated by measuring the temperature at which the genomic DNA melts. Higher temperatures are associated with high GC content.

DNA denaturation can also be used to detect sequence differences between two different DNA sequences. DNA is heated and denatured into single-stranded state, and the mixture is cooled to allow strands to rehybridize. Hybrid molecules are formed between similar sequences and any differences between those sequences will result in a disruption of the base-pairing. On a genomic scale, the method has been used by researchers to estimate the genetic distance between two species, a process known as DNA-DNA hybridization. In the context of a single isolated region of DNA, denaturing gradient gels and temperature gradient gels can be used to detect the presence of small mismatches between two sequences, a process known as temperature gradient gel electrophoresis.

Methods of DNA analysis based on melting temperature have the disadvantage of being proxies for studying the underlying sequence; DNA sequencing is generally considered a more accurate method.

The process of DNA melting is also used in molecular biology techniques, notably in the polymerase chain reaction (PCR). Although the temperature of DNA melting is not diagnostic in the technique, methods for estimating T_m are important for determining the appropriate temperatures to use in a protocol. DNA melting temperatures can also be used as a proxy for equalizing the hybridization strengths of a set of molecules, e.g. the oligonucleotide probes of DNA microarrays.

Annealing

Annealing, in genetics, means for DNA or RNA to pair by hydrogen bonds to a complementary sequence, forming a double-stranded polynucleotide. The term is often used to describe the binding of a DNA probe, or the binding of a primer to a DNA strand during a polymerase chain reaction (PCR). The term is also often used to describe the reformation (renaturation) of complementary strands that were separated by heat (thermally denatured).

Proteins such as RAD52 can help DNA anneal.

Methods for estimating melting temperatures

Several formulas are used to calculate T_m values. Some formulas are more accurate in predicting melting temperatures of DNA duplexes.

One problem in nucleic acid thermodynamics is to determine the thermodynamic parameters for forming double-stranded nucleic acid AB from single-stranded nucleic acids A and B.

$$AB \leftrightarrow A + B$$

The equilibrium constant for this reaction is $K = \frac{[A][B]}{[AB]}$. According to thermodynamics, the relation between free energy, ΔG, and K is $\Delta G^\circ = -RT\ln K$, where R is the ideal gas law constant, and T is the kelvin temperature of the reaction. This gives, for the nucleic acid system,

$$\Delta G^\circ = -RT \ln \frac{[A][B]}{[AB]}.$$

The melting temperature, T_m, occurs when half of the double-stranded nucleic acid has dissociated. If no additional nucleic acids are present, then [A], [B], and [AB] will be equal, and equal to half the initial concentration of double-stranded nucleic acid, $[AB]_{initial}$. This gives an expression for the melting point of a nucleic acid duplex of

$$T_m = \frac{-\Delta G^\circ}{R \ln \frac{[AB]_{initial}}{2}}.$$

Because $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$, T_m is also given by

$$T_m = \frac{\Delta H^\circ}{\Delta S^\circ - R \ln \frac{[AB]_{initial}}{2}}.$$

The terms ΔH° and ΔS° are usually given for the association and not the dissociation reaction. This formula then turns into :

$$T_m = \frac{\Delta H^\circ}{\Delta S^\circ + R \ln([A]_{total} - [B]_{total}/2)}$$, where $[B]_{total} < [A]_{total}$.

This equation is based on the assumption that only two states are involved in melting: the double stranded state and the random-coil state. However, nucleic acids may melt several

intermediate states. To account for such complicated behavior, the methods of statistical mechanics must be used.

Nearest-neighbor method

Some of these parameters can be determined using the nearest-neighbor method. The interaction between bases on different strands depends somewhat on the neighboring bases. Instead of treating a DNA helix as a string of interactions between base pairs, the nearest-neighbor model treats a DNA helix as a string of interactions between 'neighboring' base pairs. So, for example, the DNA shown below has nearest-neighbor interactions indicated by the arrows.

```
    ↓ ↓ ↓ ↓ ↓
5' C-G-T-T-G-A 3'
3' G-C-A-A-C-T 5'
```

The free energy of forming this DNA from the individual strands, $\Delta G°$, is represented (at 37°C) as

$\Delta G°_{37}$(predicted) = $\Delta G°_{37}$(CG initiation) + $\Delta G°_{37}$(CG/GC) + $\Delta G°_{37}$(GT/CA) + $\Delta G°_{37}$(TT/AA) + $\Delta G°_{37}$(TG/AC) + $\Delta G°_{37}$(GA/CT)

The first term represents the free energy of the first base pair, CG, in the absence of a nearest neighbor. The second term includes both the free energy of formation of the second base pair, GC, and stacking interaction between this base pair and the previous base pair. The remaining terms are similarly defined. In general, the free energy of forming a nucleic acid duplex is

$$\Delta G^\circ_{37}(total) = \Delta G^\circ_{37}(initiation) + \sum_{i=1}^{10} n_i \Delta G^\circ_{37}(i)$$

.

Each $\Delta G°$ term has enthalpic, $\Delta H°$, and entropic, $\Delta S°$, parameters, so the change in free energy is also given by

$$\Delta G^\circ(total) = \Delta H^\circ_{total} + T\Delta S^\circ_{total.}$$

Values of $\Delta H°$ and $\Delta S°$ have been determined for the ten possible pairs of interactions. These are given in Table 1, along with the value of $\Delta G°$ calculated at 37°C. Using these values, the value of $\Delta G_{37}°$ for the DNA helix shown above is calculated to be -22.4 kJ/mol. The experimental value is -21.8 kJ/mol.

Table 1. Nearest-neighbor parameters for DNA/DNA duplexes in 1 M NaCl.

Nearest-neighbor sequence (5'-3'/3'-5')	$\Delta H°$ kJ/mol	$\Delta S°$ J/(mol·K)	$\Delta G°_{37}$ kJ/(mol·K)
AA/TT	-33.1	-92.9	-4.26
AT/TA	-30.1	-85.4	-3.67
TA/AT	-30.1	-89.1	-2.50
CA/GT	-35.6	-95.0	-6.12
GT/CA	-35.1	-93.7	-6.09
CT/GA	-32.6	-87.9	-5.40
GA/CT	-34.3	-92.9	-5.51
CG/GC	-44.4	-113.8	-9.07
GC/CG	-41.0	-102.1	-9.36
GG/CC	-33.5	-83.3	-7.66
Terminal A-T base pair	9.6	17.2	4.31
Terminal G-C base pair	0.4	-11.7	4.05

The parameters associated with the ten groups of neighbors shown in table 1 are determined from melting points of short oligonucleotide duplexes. Curiously, it works out that only eight of the ten groups are independent. A more realistic way of modeling the behavior of nucleic acids would seem to be to have parameters that depend on the neighboring groups on both sides of a nucleotide, giving a table with entries like "TCG/AGC". However, this would involve around 32 groups; the number of experiments needed to get reliable data for so many groups would be considerable. Because the predictions from the nearest neighbor method agree reasonably well with experimental results, the extra effort required to develop a different model may not be justifiable.