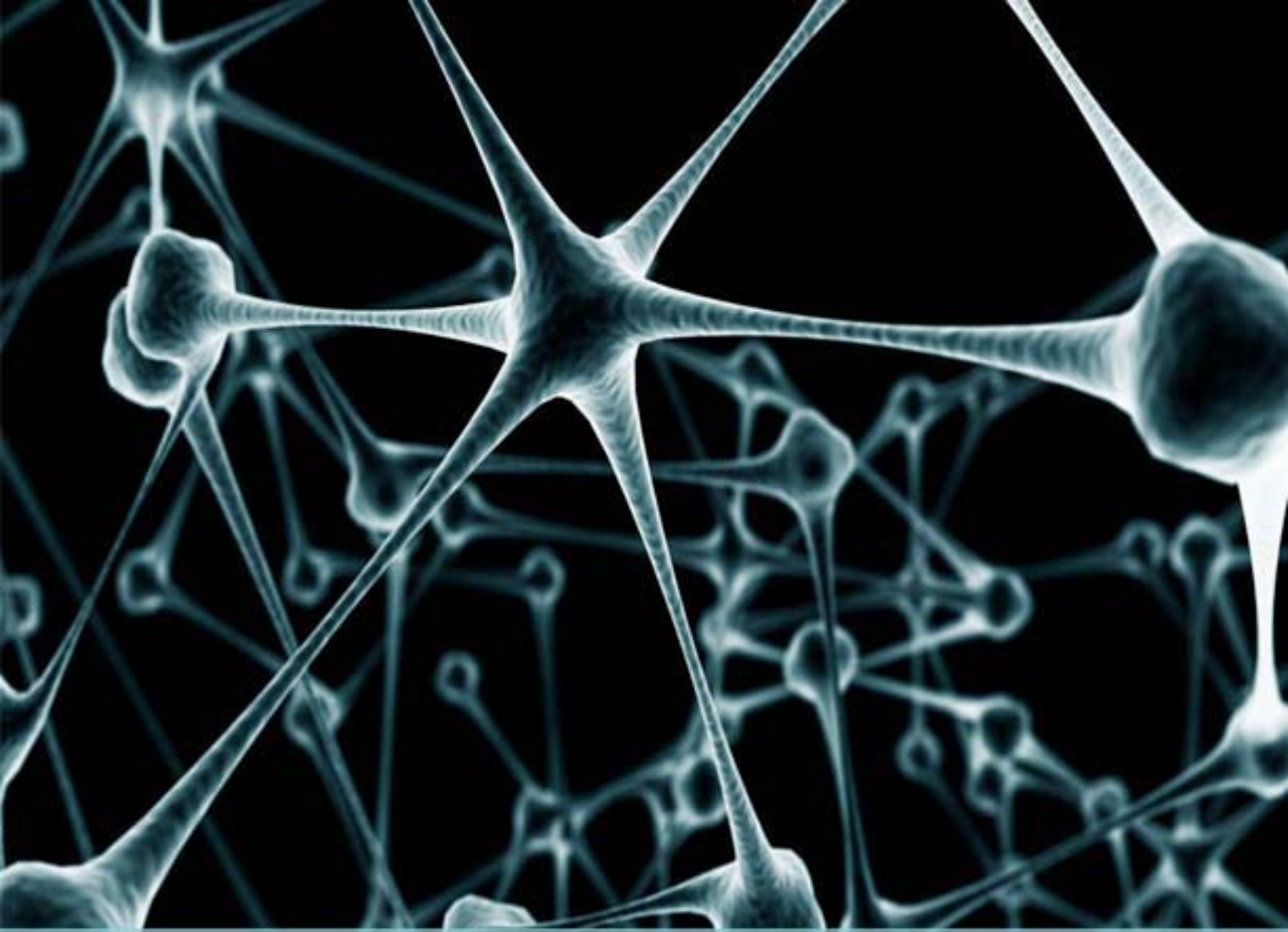


DNA Replication

Cleo Kahn

First Edition, 2012

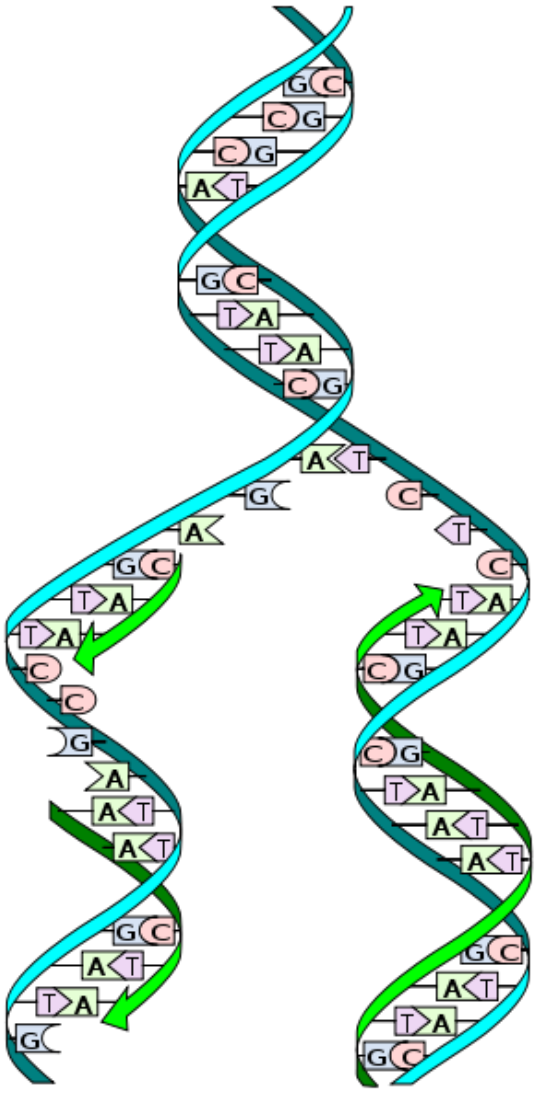
ISBN 978-81-323-3271-8

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

Table of Contents

Chapter- 1

DNA Replication



DNA replication. The double helix is unwound and each strand acts as a template. Bases are matched to synthesize the new partner strands.

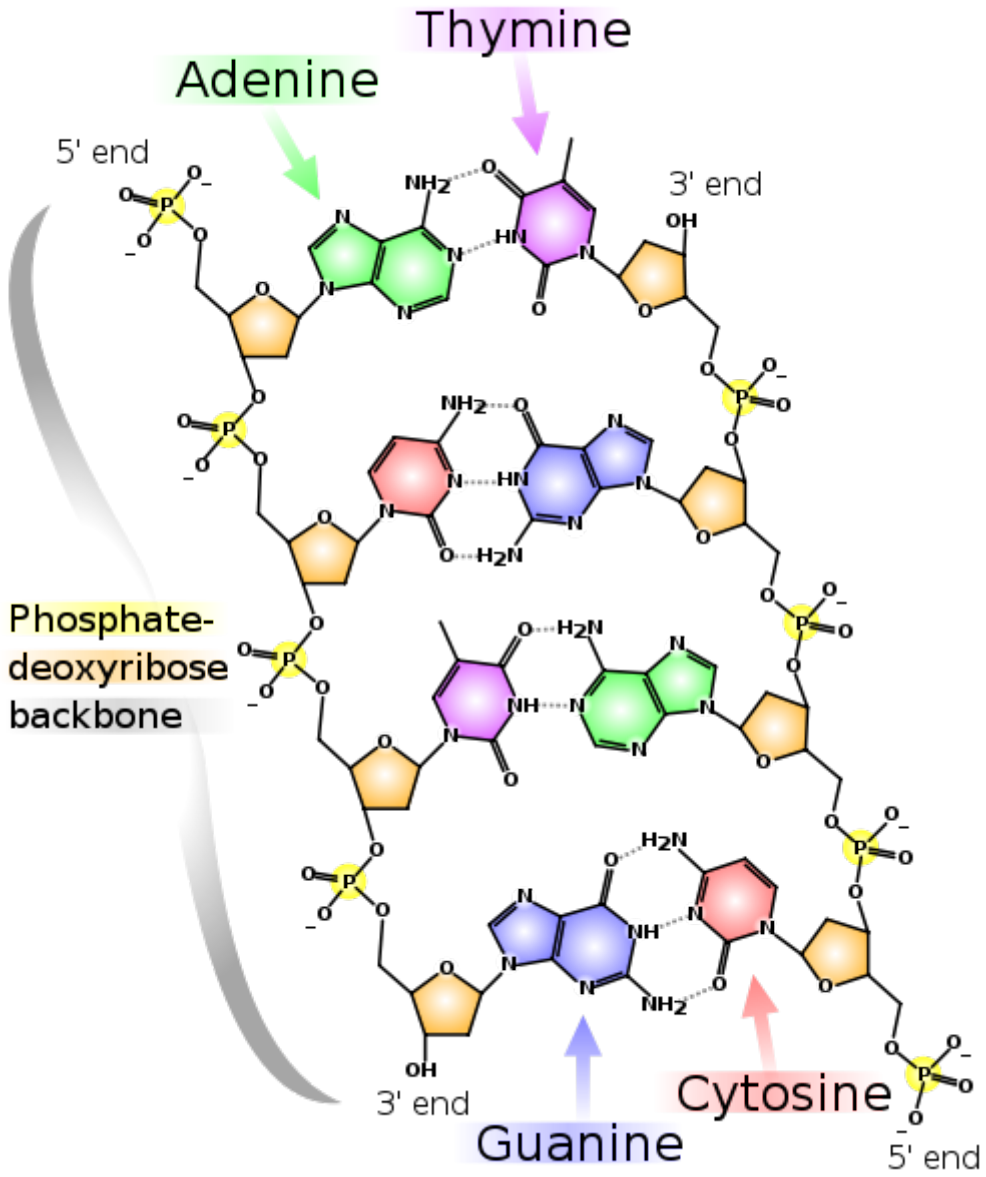
DNA replication, the basis for biological inheritance, is a fundamental process - that occurs in all living organisms - that copies their DNA. This process is *replication*, in that

each strand of the original double-stranded DNA molecule serves as template for the reproduction of the complementary strand. Therefore, following DNA replication, two identical DNA molecules have been produced from a single double-stranded DNA molecule. Cellular proofreading and error toe-checking mechanisms ensure near perfect fidelity for DNA replication.

In a cell, DNA replication begins at specific locations in the genome, called "origins". Unwinding of DNA at the origin, and synthesis of new strands, forms a replication fork. In addition to DNA polymerase, the enzyme that synthesizes the new DNA by adding nucleotides matched to the template strand, a number of other proteins are associated with the fork and assist in the initiation and continuation of DNA synthesis.

DNA replication can also be performed *in vitro* (outside a cell). DNA polymerases, isolated from cells, and artificial DNA primers are used to initiate DNA synthesis at known sequences in a template molecule. The polymerase chain reaction (PCR), a common laboratory technique, employs such artificial synthesis in a cyclic manner to amplify a specific target DNA fragment from a pool of DNA.

DNA structure



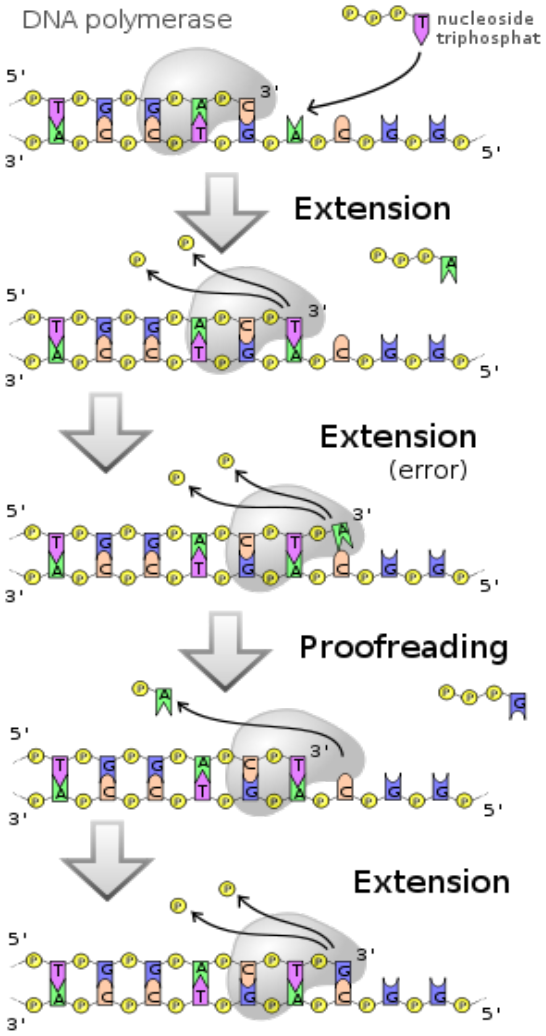
The chemical structure of DNA

DNA usually exists as a double-stranded structure, with both strands coiled together to form the characteristic double-helix. Each single strand of DNA is a chain of four types of nucleotides having the bases: adenine, cytosine, guanine, and thymine. A nucleotide is a mono-, di-, or triphosphate deoxyribonucleoside; that is, a deoxyribose sugar is attached to one, two, or three phosphates. Chemical interaction of these nucleotides forms phosphodiester linkages, creating the phosphate-deoxyribose backbone of the DNA double helix with the bases pointing inward. Nucleotides (bases) are matched between strands through hydrogen bonds to form base pairs. Adenine pairs with thymine and cytosine pairs with guanine.

DNA strands have a directionality, and the different ends of a single strand are called the "3' (three-prime) end" and the "5' (five-prime) end." These terms refer to the carbon atom in deoxyribose to which the next phosphate in the chain attaches. In addition to being complementary, the two strands of DNA are antiparallel: They are orientated in opposite directions. This directionality has consequences in DNA synthesis, because DNA polymerase can synthesize DNA in only one direction by adding nucleotides to the 3' end of a DNA strand.

The pairing of bases in DNA through hydrogen bonding means that the information contained within each strand is redundant. The nucleotides on a single strand can be used to reconstruct nucleotides on a newly synthesized partner strand.

DNA polymerase



DNA polymerases adds nucleotides to the 5' end of a strand of DNA. If a mismatch is accidentally incorporated, the polymerase is inhibited from further extension. Proofreading removes the mismatched nucleotide and extension continues.

DNA polymerases are a family of enzymes that carry out all forms of DNA replication. A DNA polymerase can only extend an existing DNA strand paired with a template strand; it cannot begin the synthesis of a new strand. To begin synthesis of a new strand, a short fragment of DNA or RNA, called a primer, must be created and paired with the template strand before DNA polymerase can synthesize new DNA.

Once a primer pairs with DNA to be replicated, DNA polymerase synthesizes a new strand of DNA by extending the 3' end of an existing nucleotide chain, adding new nucleotides matched to the template strand one at a time via the creation of phosphodiester bonds. The energy for this process of DNA polymerization comes from two of the three total phosphates attached to each unincorporated base. (Free bases with their attached phosphate groups are called nucleoside triphosphates.) When a nucleotide is being added to a growing DNA strand, two of the phosphates are removed and the energy produced creates a phosphodiester (chemical) bond that attaches the remaining phosphate to the growing chain. The energetics of this process also help explain the directionality of synthesis - if DNA were synthesized in the 3' to 5' direction, the energy for the process would come from the 5' end of the growing strand rather than from free nucleotides.

In general, DNA polymerases are extremely accurate, making less than one error for every 10^7 nucleotides added. Even so, some DNA polymerases also have proofreading ability; they can remove nucleotides from the end of a strand in order to correct mismatched bases. If the 5' nucleotide needs to be removed during proofreading, the triphosphate end is lost. Hence, the energy source that usually provides energy to add a new nucleotide is also lost.

DNA replication within the cell

Origins of replication

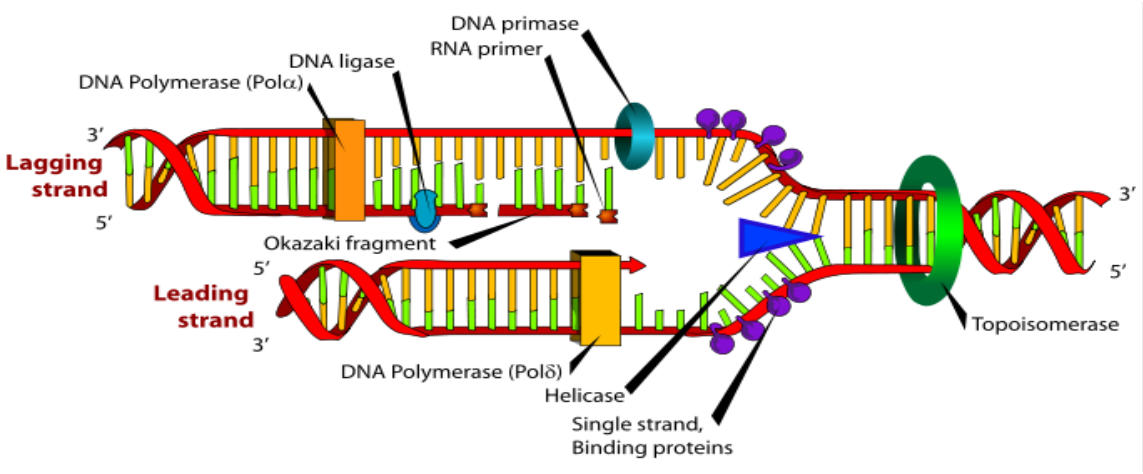
For a cell to divide, it must first replicate its DNA. This process is initiated at particular points within the DNA, known as "origins", which are targeted by proteins that separate the two strands and initiate DNA synthesis. Origins contain DNA sequences recognized by replication initiator proteins (e.g., dnaA in *E coli'* and the Origin Recognition Complex in yeast). These initiator proteins recruit other proteins to separate the two strands and initiate replication forks.

Initiator proteins recruit other proteins to separate the DNA strands at the origin, forming a bubble. Origins tend to be "AT-rich" (rich in adenine and thymine bases) to assist this process, because A-T base pairs have two hydrogen bonds (rather than the three formed in a C-G pair)—in general, strands rich in these nucleotides are easier to separate due the positive relationship between the number of hydrogen bonds and the difficulty of breaking these bonds. Once strands are separated, RNA primers are created on the template strands. To be more specific, the leading strand receives one RNA primer per active origin of replication while the lagging strand receives several; these several fragments of RNA primers found on the lagging strand of DNA are called Okazaki

fragments, named after their discoverer. DNA Polymerase extends the leading strand in one continuous motion and the lagging strand in a discontinuous motion (due to the Okazaki fragments). RNase removes the RNA fragments used to initiate replication by DNA Polymerase, and another DNA Polymerase enters to fill the gaps. When this is complete, a single nick on the leading strand and several nicks on the lagging strand can be found. Ligase works to fill these nicks in, thus completing the newly replicated DNA molecule.

As DNA synthesis continues, the original DNA strands continue to unwind on each side of the bubble, forming 2 replication forks. In bacteria, which have a single origin of replication on their circular chromosome, this process eventually creates a "theta structure" (resembling the Greek letter theta: θ). In contrast, eukaryotes have longer linear chromosomes and initiate replication at multiple origins within these.

The replication fork



Many enzymes are involved in the DNA replication fork.

The replication fork is a structure that forms within the nucleus during DNA replication. It is created by helicases, which break the hydrogen bonds holding the two DNA strands together. The resulting structure has two branching "prongs", each one made up of a single strand of DNA. These two strands serve as the template for the leading and lagging strands, which will be created as DNA polymerase matches complementary nucleotides to the templates; The templates may be properly referred to as the leading strand template and the lagging strand template.

Leading strand

The leading strand is the template strand of the DNA double helix so that the replication fork moves along it in the 3' to 5' direction. This allows the new strand synthesized complementary to it to be synthesized 5' to 3' in the same direction as the movement of the replication fork.

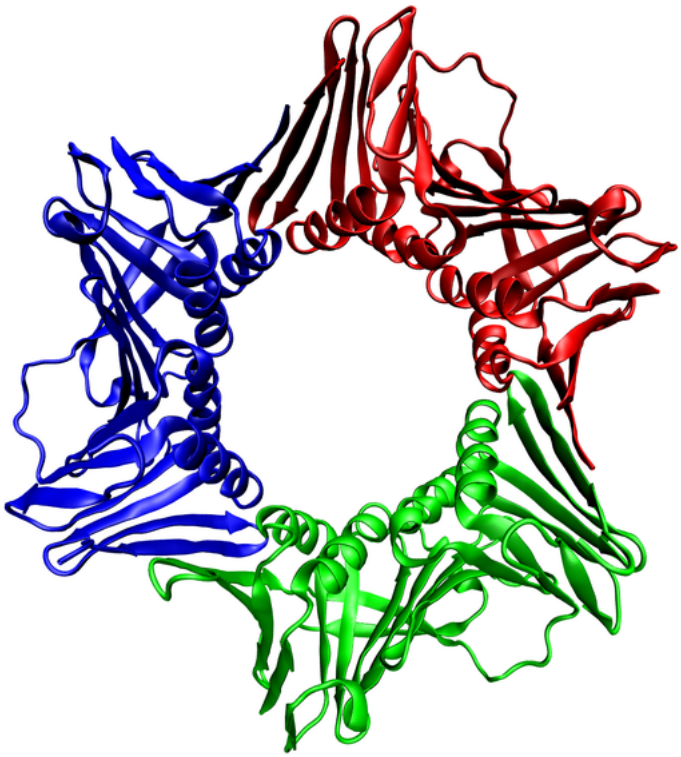
On the leading strand, a polymerase "reads" the DNA and adds nucleotides to it continuously. This polymerase is DNA polymerase III (DNA Pol III) in prokaryotes and presumably Pol ε in eukaryotes.

Lagging strand

The lagging strand is the strand of the template DNA double helix that is oriented so that the replication fork moves along it in a 5' to 3' manner. Because of its orientation, opposite to the working orientation of DNA polymerase III, which moves on a template in a 3' to 5' manner, replication of the lagging strand is more complicated than that of the leading strand.

On the lagging strand, primase "reads" the DNA and adds RNA to it in short, separated segments. In eukaryotes, primase is intrinsic to Pol α. DNA polymerase III or Pol δ lengthens the primed segments, forming Okazaki fragments. Primer removal in eukaryotes is also performed by Pol δ. In prokaryotes, DNA polymerase I "reads" the fragments, removes the RNA using its flap endonuclease domain (RNA primers are removed by 5'-3' exonuclease activity of polymerase I [weaver, 2005], and replaces the RNA nucleotides with DNA nucleotides (this is necessary because RNA and DNA use slightly different kinds of nucleotides). DNA ligase joins the fragments together.

Dynamics at the replication fork



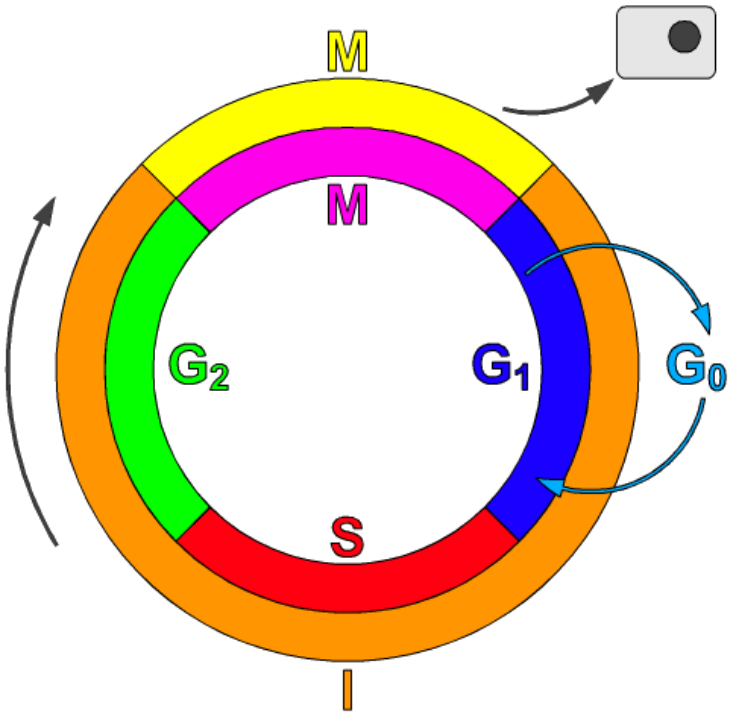
The assembled human DNA clamp, a trimer of the protein PCNA

As helicase unwinds DNA at the replication fork, the DNA ahead is forced to rotate. This process results in a build-up of twists in the DNA ahead. This build-up would form a resistance that would eventually halt the progress of the replication fork. DNA topoisomerases are enzymes that solve these physical problems in the coiling of DNA. Topoisomerase I cuts a single backbone on the DNA, enabling the strands to swivel around each other to remove the build-up of twists. Topoisomerase II cuts both backbones, enabling one double-stranded DNA to pass through another, thereby removing knots and entanglements that can form within and between DNA molecules.

Bare single-stranded DNA has a tendency to fold back upon itself and form secondary structures; these structures can interfere with the movement of DNA polymerase. To prevent this, single-strand binding proteins bind to the DNA until a second strand is synthesized, preventing secondary structure formation.

Clamp proteins form a sliding clamp around DNA, helping the DNA polymerase maintain contact with its template, thereby assisting with processivity. The inner face of the clamp enables DNA to be threaded through it. Once the polymerase reaches the end of the template or detects double-stranded DNA, the sliding clamp undergoes a conformational change that releases the DNA polymerase. Clamp-loading proteins are used to initially load the clamp, recognizing the junction between template and RNA primers.

Regulation of replication



The cell cycle of eukaryotic cells

Eukaryotes

Within eukaryotes, DNA replication is controlled within the context of the cell cycle. As the cell grows and divides, it progresses through stages in the cell cycle; DNA replication occurs during the S phase (**S**ynthesis phase). The progress of the eukaryotic cell through the cycle is controlled by cell cycle checkpoints. Progression through checkpoints is controlled through complex interactions between various proteins, including cyclins and cyclin-dependent kinases.

The G1/S checkpoint (or restriction checkpoint) regulates whether eukaryotic cells enter the process of DNA replication and subsequent division. Cells that do not proceed through this checkpoint are quiescent in the "G0" stage and do not replicate their DNA.

Replication of chloroplast and mitochondrial genomes occurs independent of the cell cycle, through the process of D-loop replication.

Bacteria

Most bacteria do not go through a well-defined cell cycle and, instead, continuously copy their DNA; during rapid growth, this can result in the concurrent occurrences of multiple rounds of replication. Within *E coli*, the best-characterized bacteria, regulation of DNA replication can be achieved through several mechanisms, including: the hemimethylation and sequestering of the origin sequence, the ratio of ATP to ADP, and the levels of protein DnaA. These all control the process of initiator proteins binding to the origin sequences.

Because *E coli* methylates GATC DNA sequences, DNA synthesis results in hemimethylated sequences. This hemimethylated DNA is recognized by a protein (SeqA), which binds and sequesters the origin sequence; in addition, dnaA (required for initiation of replication) binds less well to hemimethylated DNA. As a result, newly replicated origins are prevented from immediately initiating another round of DNA replication.

ATP builds up when the cell is in a rich medium, triggering DNA replication once the cell has reached a specific size. ATP competes with ADP to bind to DnaA, and the DnaA-ATP complex is able to initiate replication. A certain number of DnaA proteins are also required for DNA replication — each time the origin is copied the number of binding sites for DnaA doubles, requiring the synthesis of more DnaA to enable another initiation of replication.

Termination of replication

Because bacteria have circular chromosomes, termination of replication occurs when the two replication forks meet each other on the opposite end of the parental chromosome. *E coli* regulate this process through the use of termination sequences that, when bound by the Tus protein, enable only one direction of replication fork to pass through. As a result,

the replication forks are constrained to always meet within the termination region of the chromosome.

Eukaryotes initiate DNA replication at multiple points in the chromosome, so replication forks meet and terminate at many points in the chromosome; these are not known to be regulated in any particular manner. Because eukaryotes have linear chromosomes, DNA replication is unable to synthesize to the very end of the chromosomes (telomeres), resulting in telomere shortening. This is a normal process in somatic cells — cells are able to divide only a certain number of times before the DNA loss prevents further division. (This is known as the Hayflick limit.) Within the germ cell line, which passes DNA to the next generation, telomerase extends the repetitive sequences of the telomere region to prevent degradation. Telomerase can become mistakenly active in somatic cells, sometimes leading to cancer formation.

Polymerase chain reaction

Researchers commonly replicate DNA *in vitro* using the polymerase chain reaction (PCR). PCR uses a pair of primers to span a target region in template DNA, and then polymerizes partner strands in each direction from these primers using a thermostable DNA polymerase. Repeating this process through multiple cycles produces amplification of the targeted DNA region. At the start of each cycle, the mixture of template and primers is heated, separating the newly synthesized molecule and template. Then, as the mixture cools, both of these become templates for annealing of new primers, and the polymerase extends from these. As a result, the number of copies of the target region doubles each round, increasing exponentially.

Chapter- 2

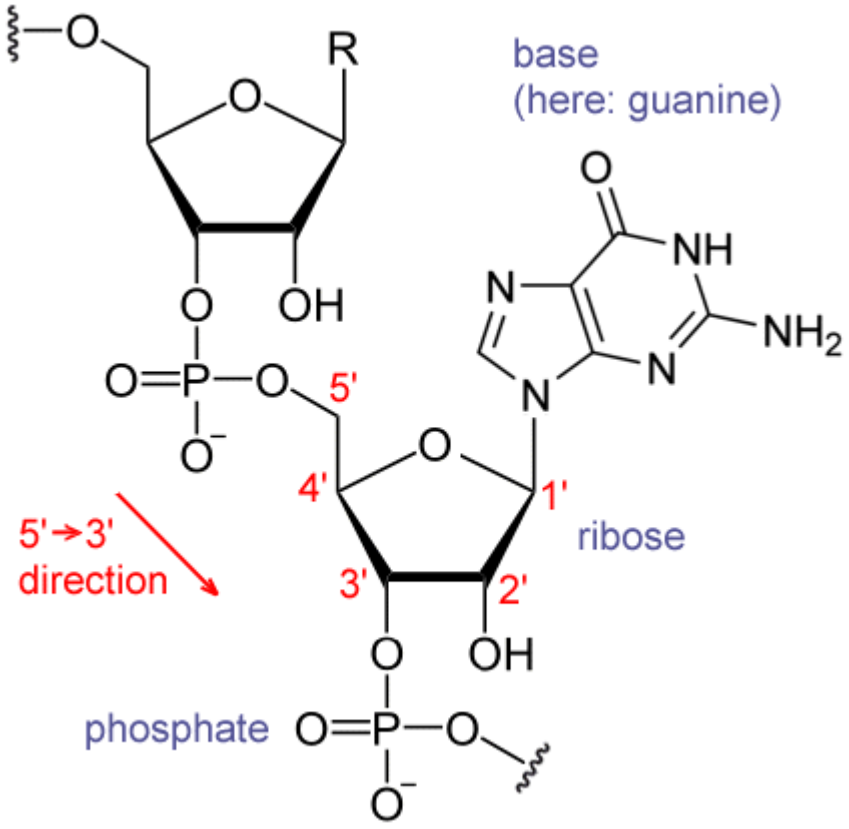
Nucleic Acid Sequence

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

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

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

Nucleotides



Chemical structure of RNA

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

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

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

Notation

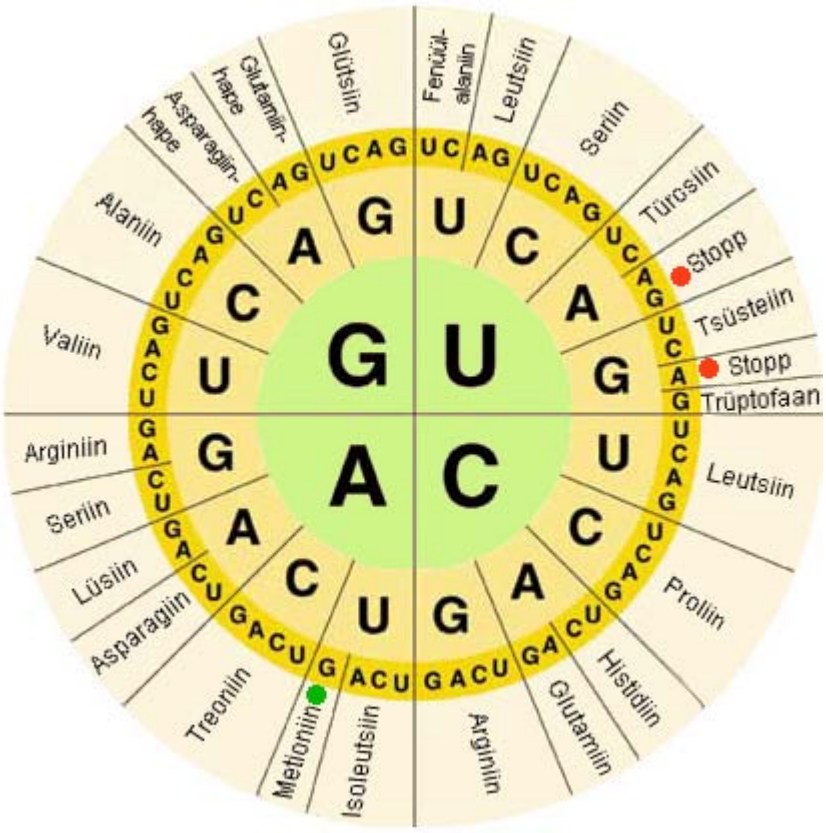
While A, T, C, and G represent a particular nucleotide at a position, there are also letters that represent ambiguity. Of all the molecules sampled, there is more than one kind of nucleotide at that position. The rules of the International Union of Pure and Applied Chemistry (IUPAC) are as follows:

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

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

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

Biological significance

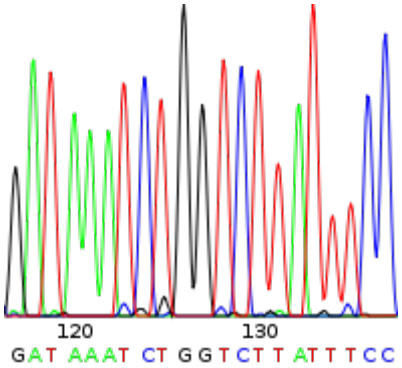


A depiction of the genetic code, by which the information contained in nucleic acids are translated into amino acid sequences in proteins.

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

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

Sequence determination



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

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

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

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

Digital format

Genetic sequence in digital format

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

Sequence analysis

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

Genetic testing

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

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

Sequence alignment

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

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

Sequence motifs

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

Chapter- 3

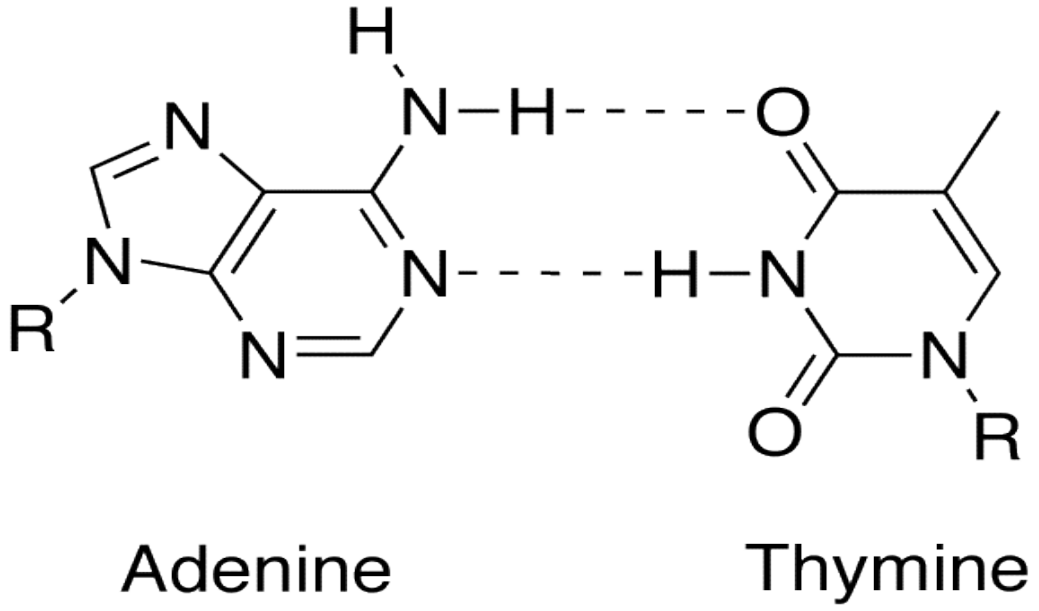
Nucleic Acid Secondary Structure

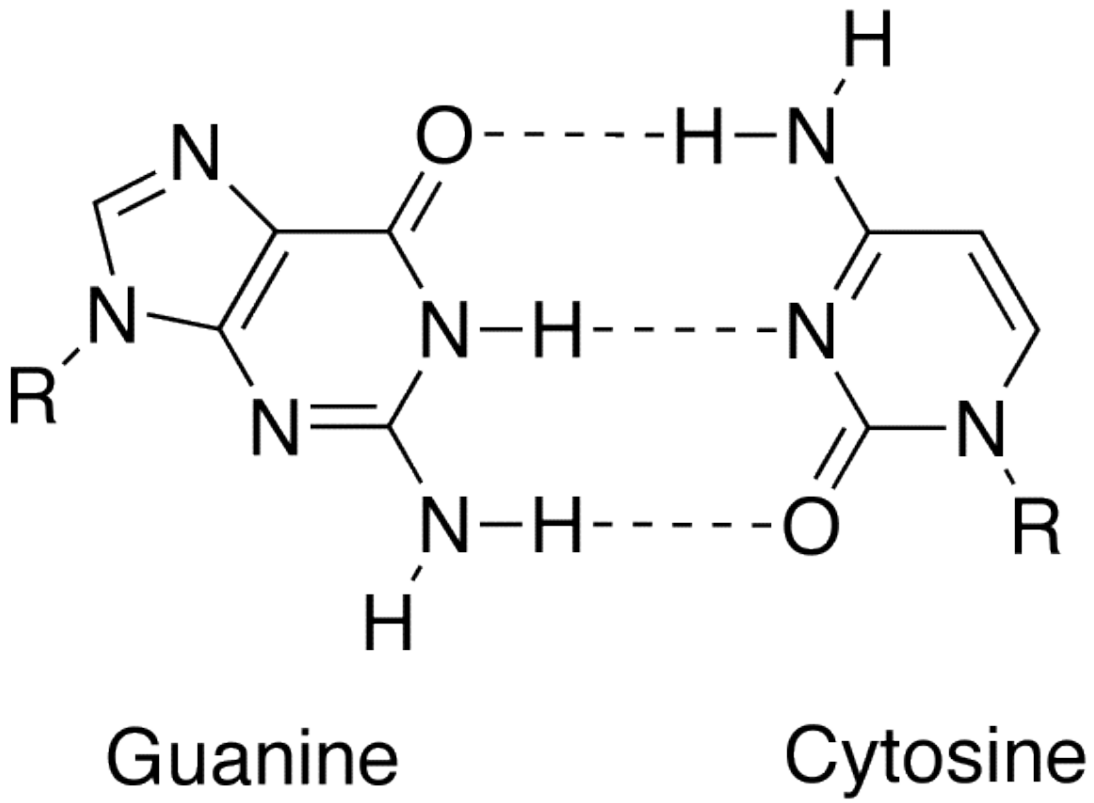
The **secondary structure of a nucleic acid molecule** refers to the basepairing interactions within a single molecule or set of interacting molecules, and can be represented as a list of bases which are paired in a nucleic acid molecule. The secondary structures of biological DNA's and RNA's tend to 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





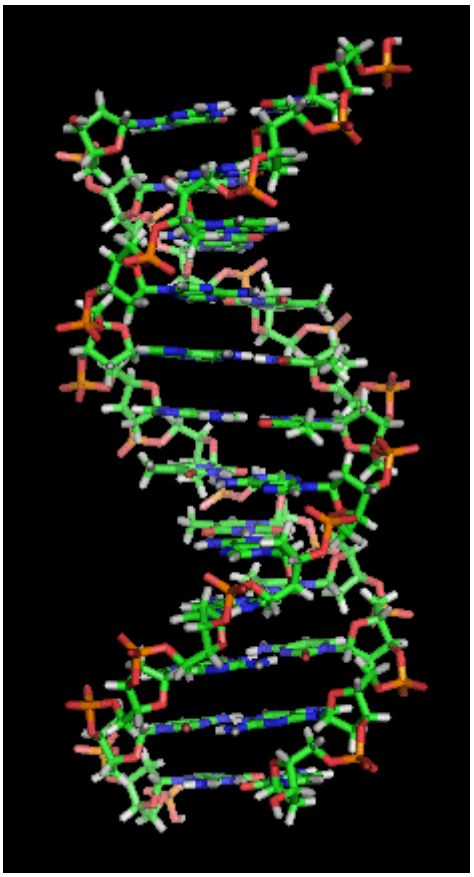
Top, an AT base pair demonstrating two intermolecular hydrogen bonds; *bottom*, a GC base pair demonstrating three intermolecular hydrogen bonds.

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

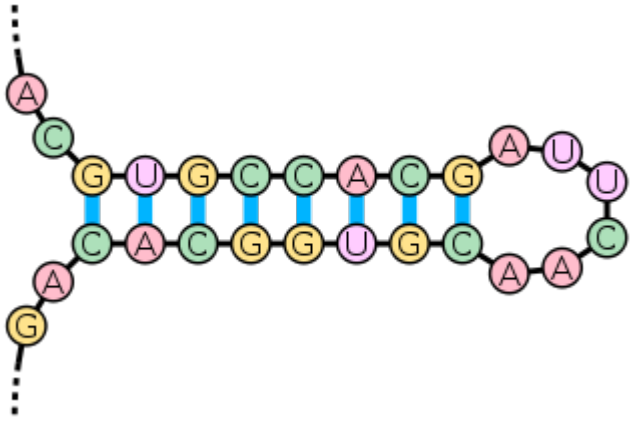
Nucleic acid secondary structure is generally divided into helices (contiguous base pairs), and various kinds of loops (unpaired nucleotides surrounded by helices). Frequently these elements, or combinations of them, can be further classified, for example, tetraloops, pseudoknots, and stem-loops.

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

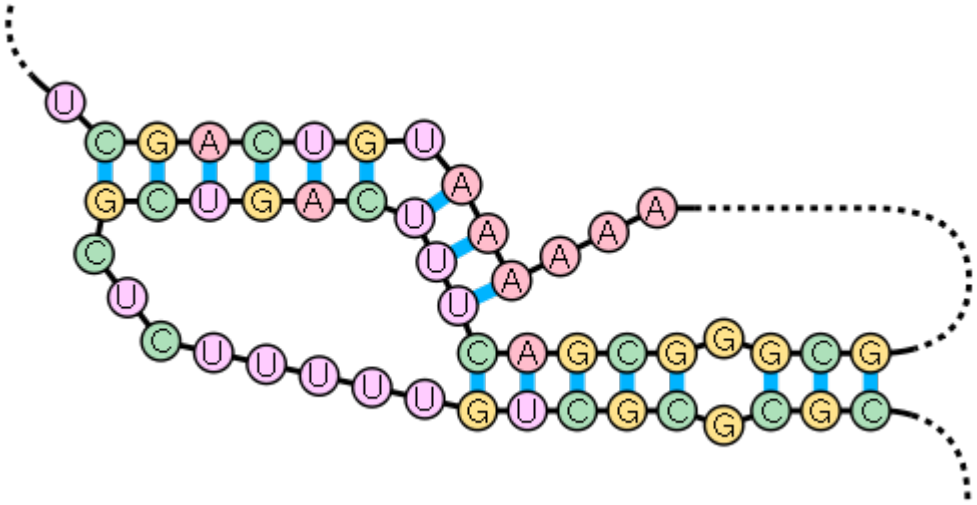


An example of an RNA stem-loop secondary structure

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.

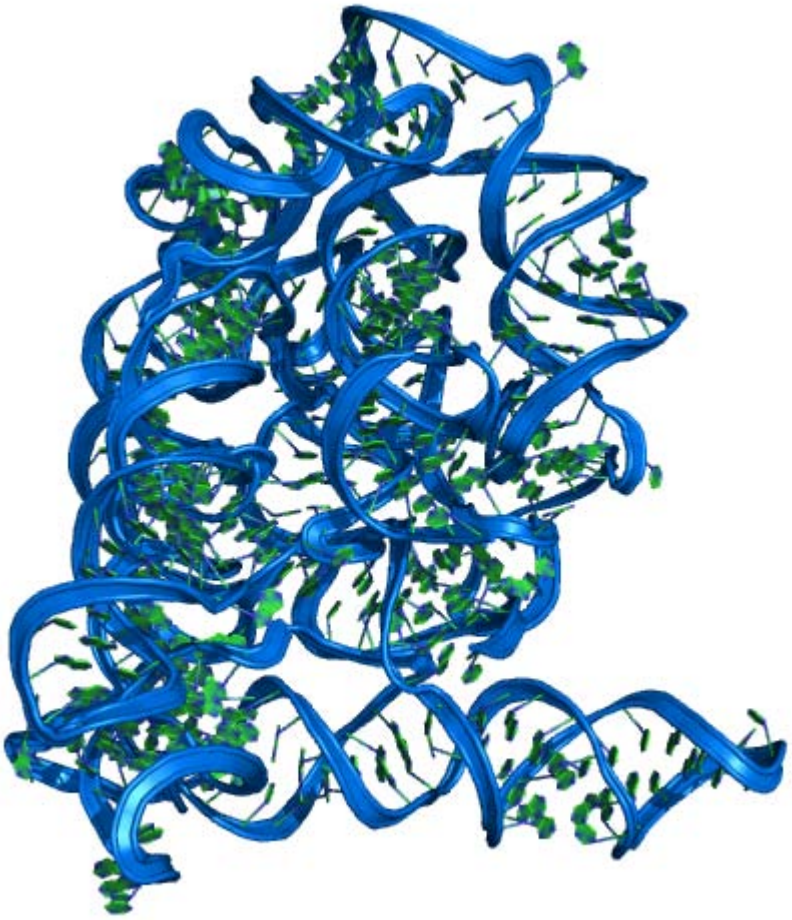
Secondary structure prediction

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- 4

Nucleic Acid Tertiary Structure

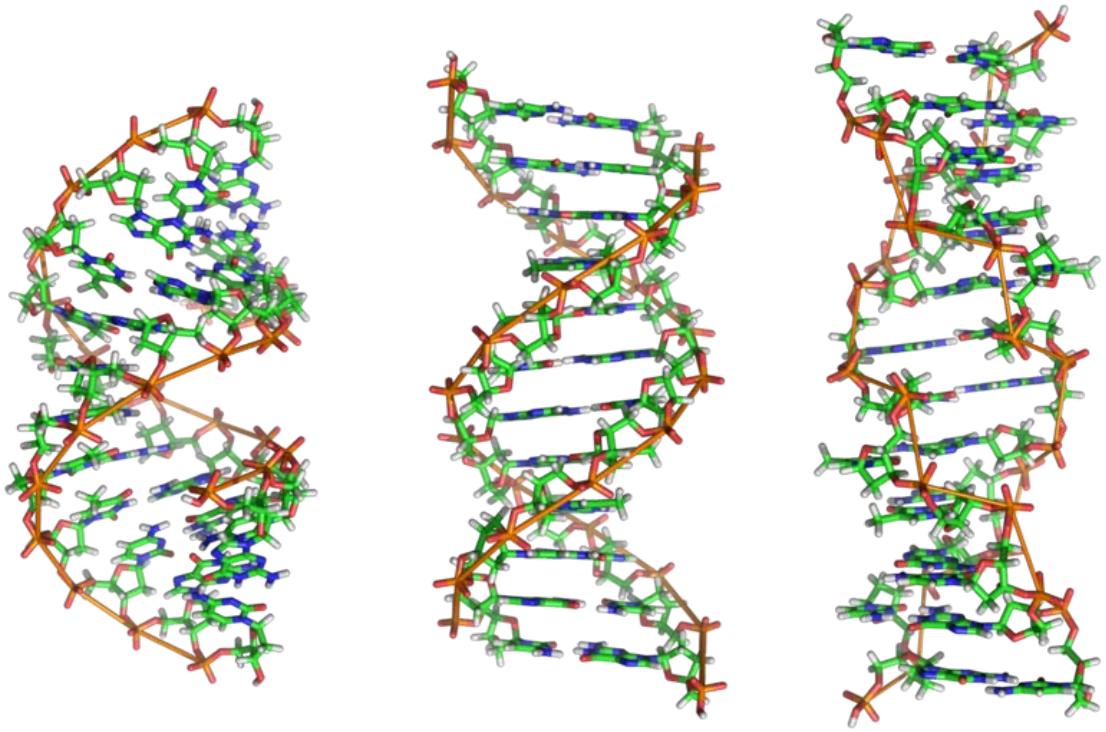


Example of a large catalytic RNA. The self-splicing group II intron from *Oceanobacillus iheyensis*.

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



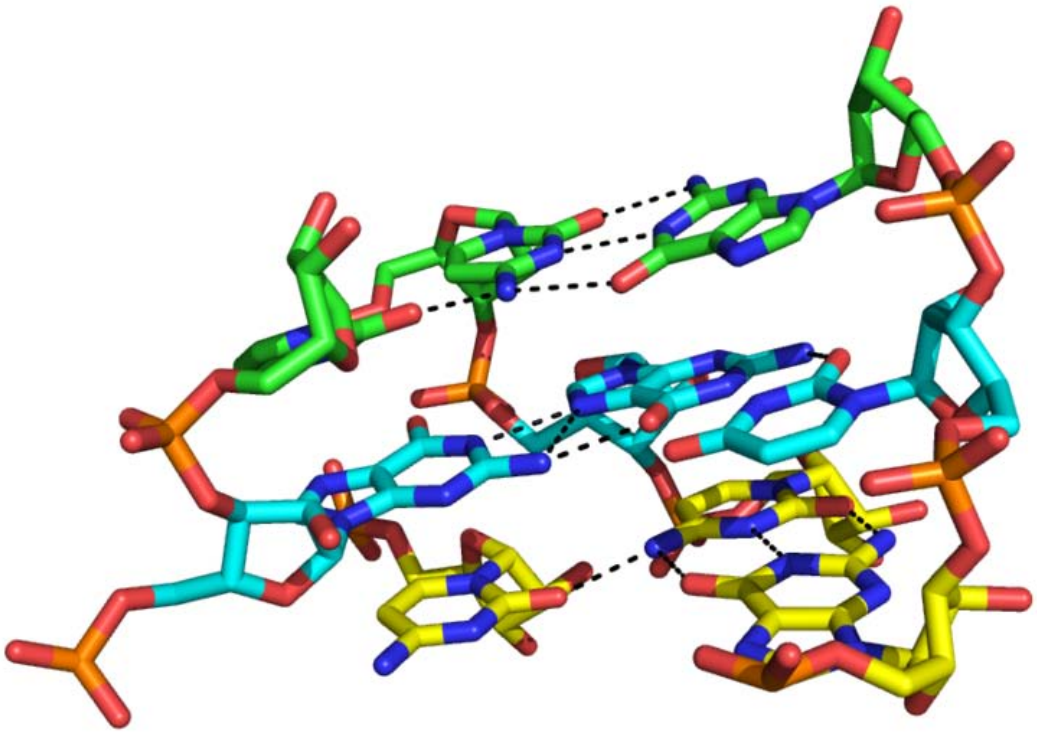
The structures of the A-, B-, and Z-DNA double helix structures

Double helix

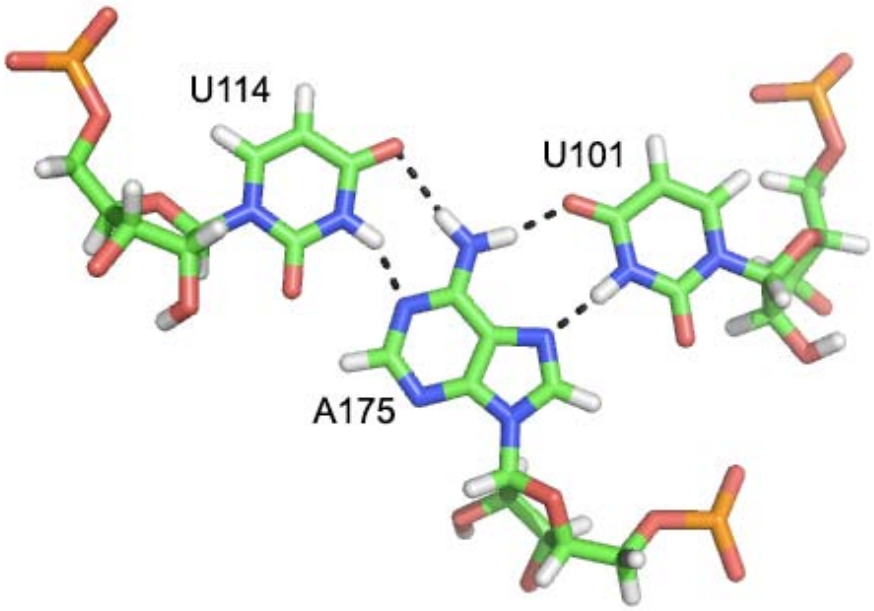
The double helix is the dominant tertiary structure for biological DNA, and is also a possible structure for RNA. Three DNA conformations are believed to be found in nature, A-DNA, B-DNA, and Z-DNA. The "B" form described by James D. Watson and Francis Crick is believed to predominate in cells. James 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.



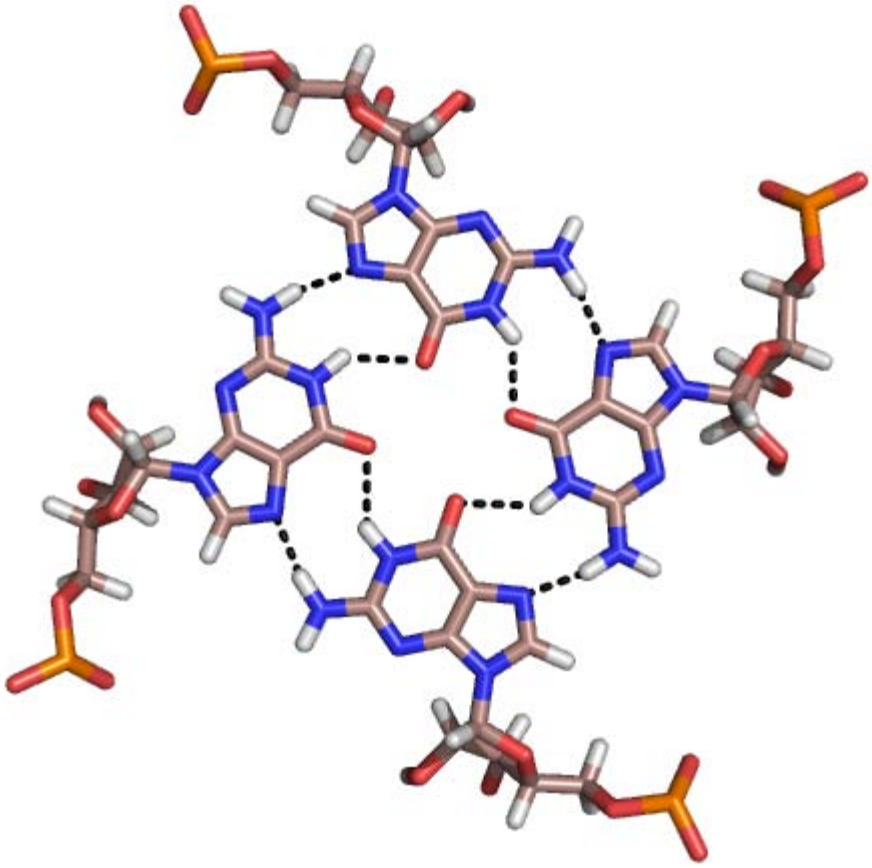
Close-up rendering of the U114:A175-U101 major groove (Hoogsteen base) triplex formed within the wild type pseudoknot of Human Telomerase RNA. Hydrogen bonds are shown in black dashed lines. "N" atoms are colored in blue and "o" atoms in red.

Major and minor groove triplexes

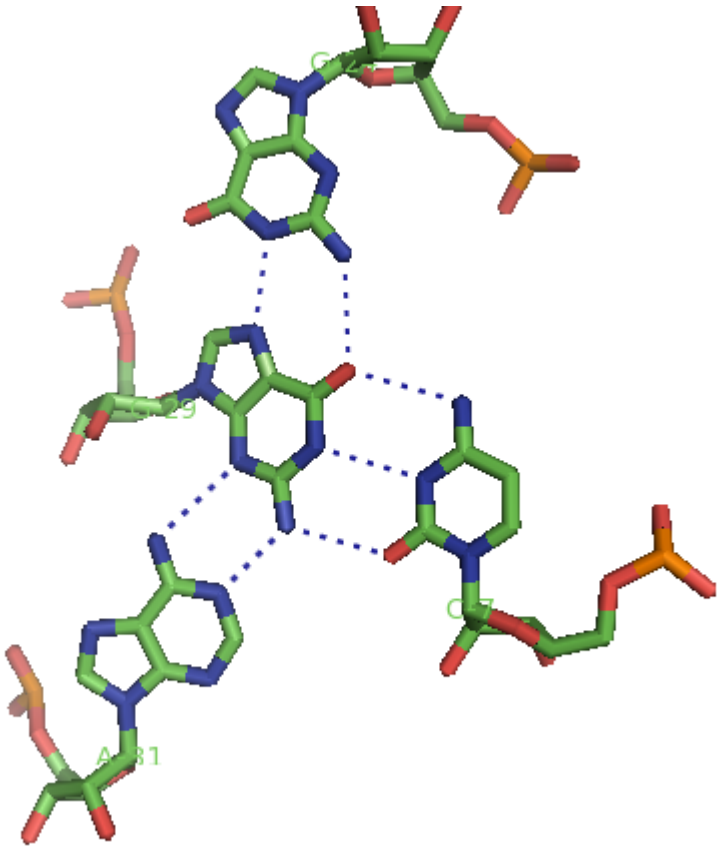
The minor groove triple is a ubiquitous RNA structural motif. Because interactions with the minor groove are often mediated by the 2'-OH of the ribose sugar, this RNA motif looks very different from its DNA equivalent. The most common example of a minor loop triple is the A-minor motif, or the insertion of adenosine bases into the minor groove (see above). 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

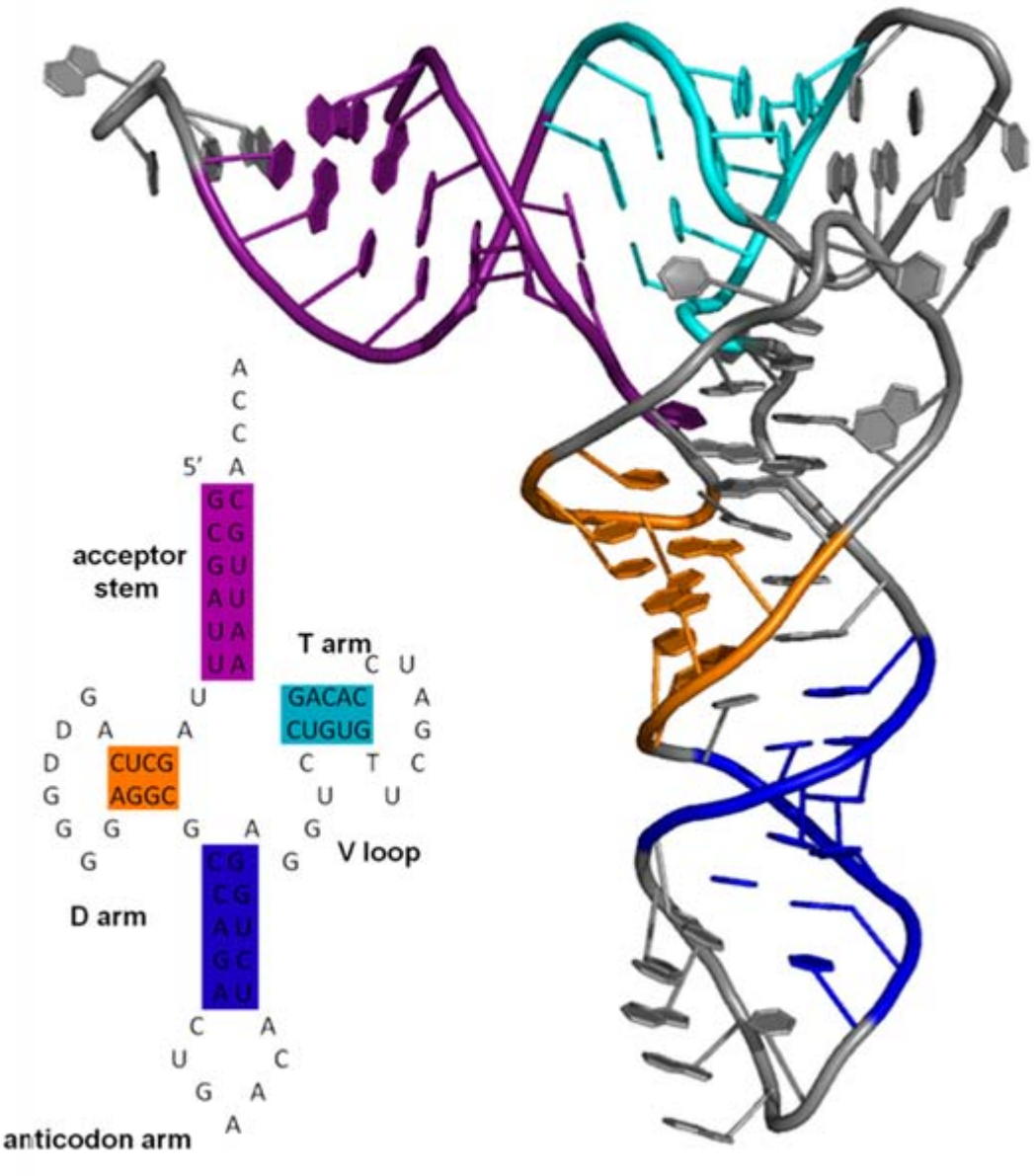
Besides double helices and the above-mentioned triplexes, RNA and DNA can both also form quadruple helices. There are diverse structures of RNA base quadruplexes. Four consecutive guanine residues can form a quadruplex in RNA by Hoogsteen hydrogen bonds to form a “Hoogsteen ring” (See Figure). G-C and 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 (See Figure). 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.

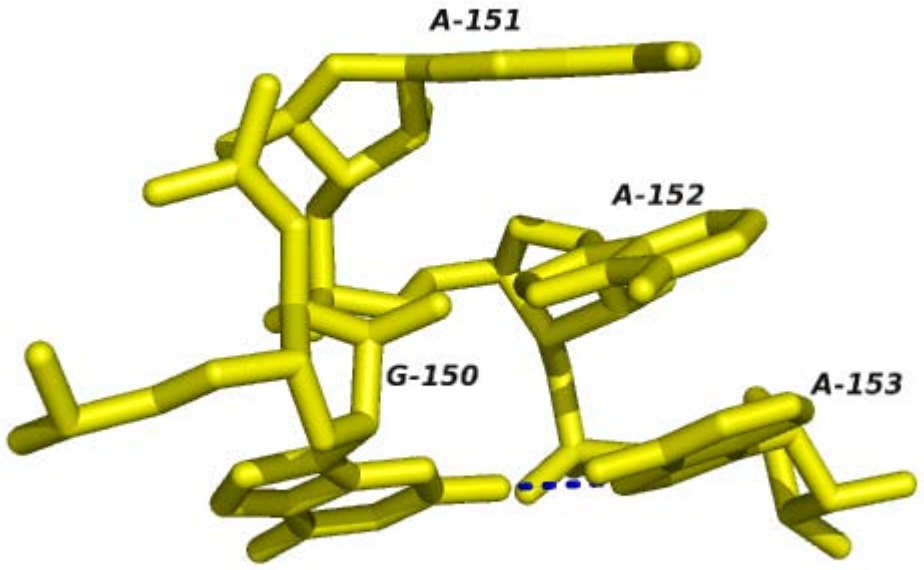
Most well-studied RNA tertiary structures contain examples of coaxial stacking. Some prominent examples are tRNA-Phe, group I introns, group II introns, and ribosomal RNAs. Crystal structures of tRNA revealed the presence of two extended helices that result from coaxial stacking of the amino-acid acceptor 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

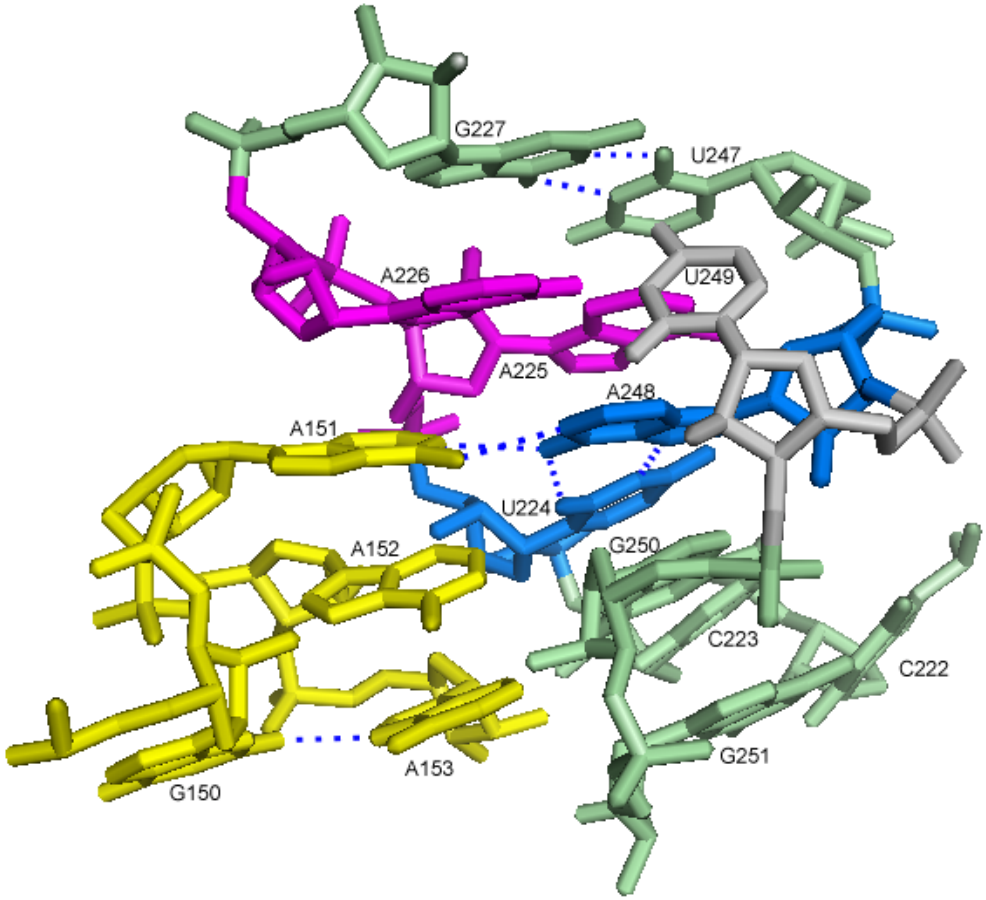
Tetraloop-receptor interactions



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.



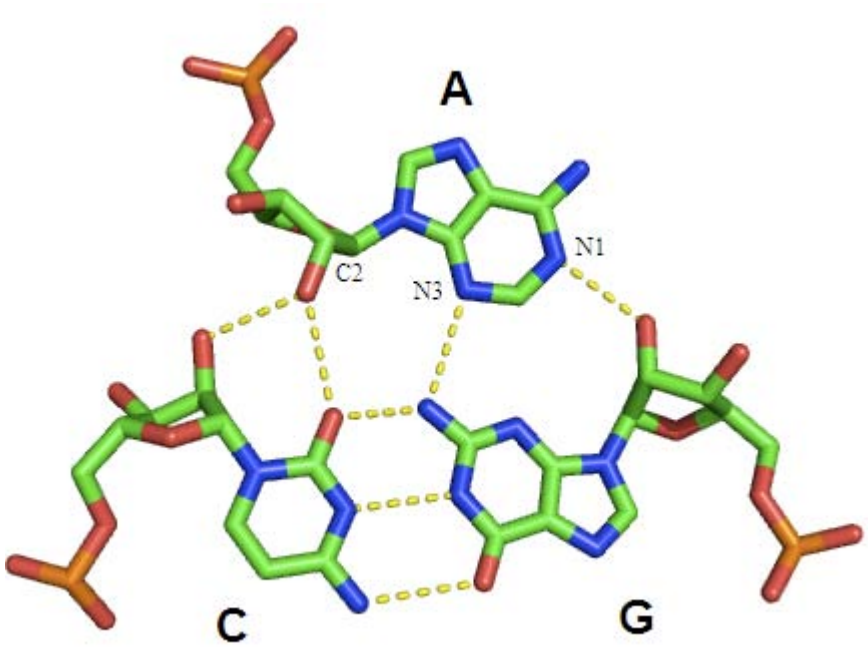
GAAA Tetraloop and Receptor: Stick representation of tetraloop (yellow) and its receptor, showing both Watson-Crick and Hoogsteen base-pairing.

"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 (see above) 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.

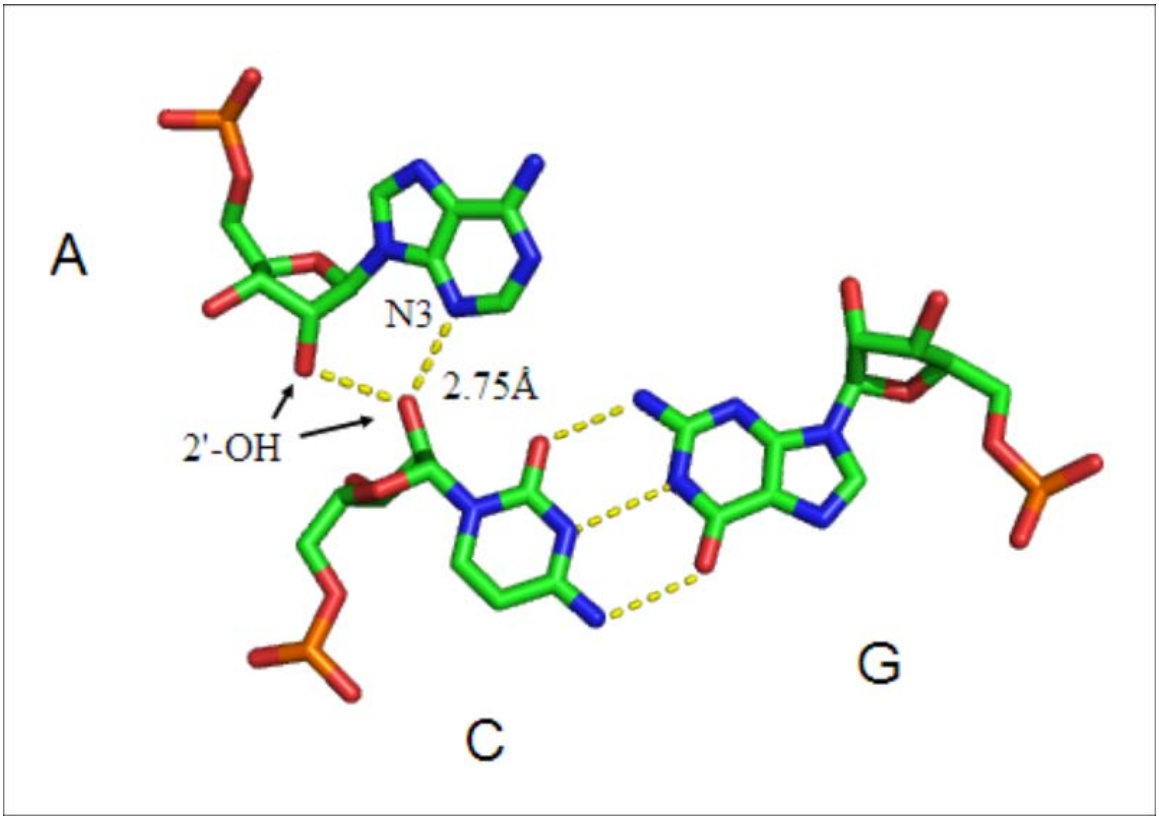
For example, the self-splicing group I intron relies on tetraloop receptor motifs for its structure and function. Specifically, the three adenine residues of the canonical GAAA motif stack on top of the receptor helix and form multiple stabilizing hydrogen bonds with the receptor. The first adenine of 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.

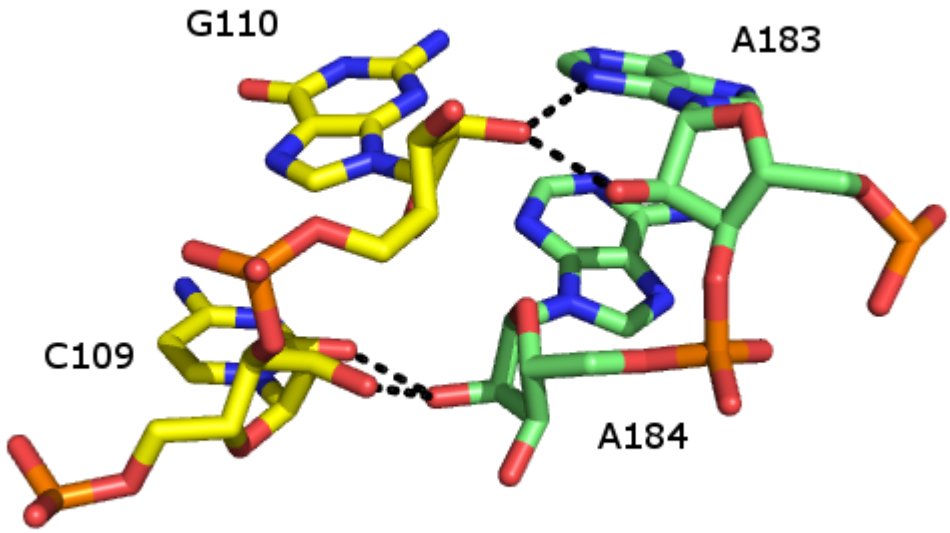
The A-minor motif is a ubiquitous RNA tertiary structural motif. It is formed by the insertion of an unpaired nucleoside into the minor groove of an RNA duplex. As such it is an example of a minor groove triple. Although guanosine, cytosine and uridine can also form minor groove triple interactions, 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.

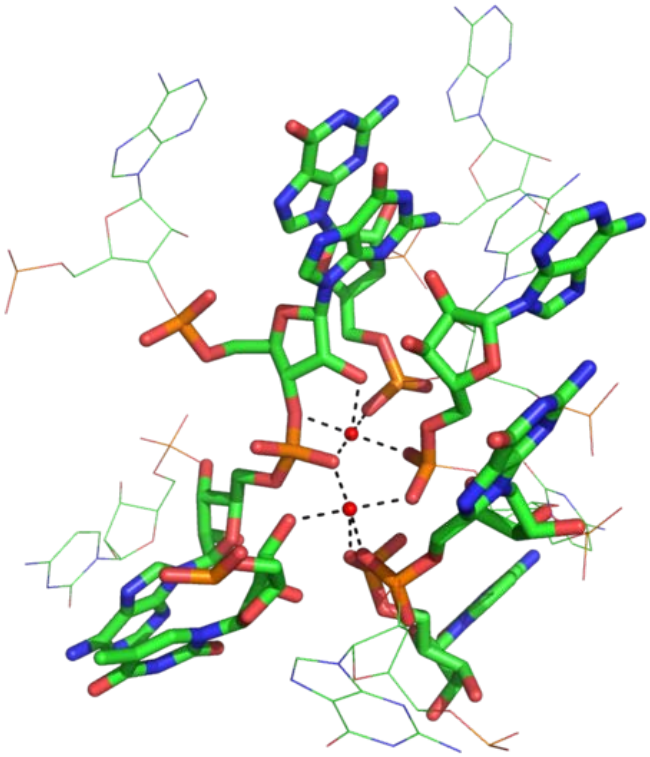
The ribose zipper is an RNA tertiary structural element in which two RNA chains are held together by hydrogen bonding interactions involving the 2'OH of ribose sugars on different strands. The 2'OH can behave as both hydrogen bond donor and acceptor, which allows formation of bifurcated hydrogen bonds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

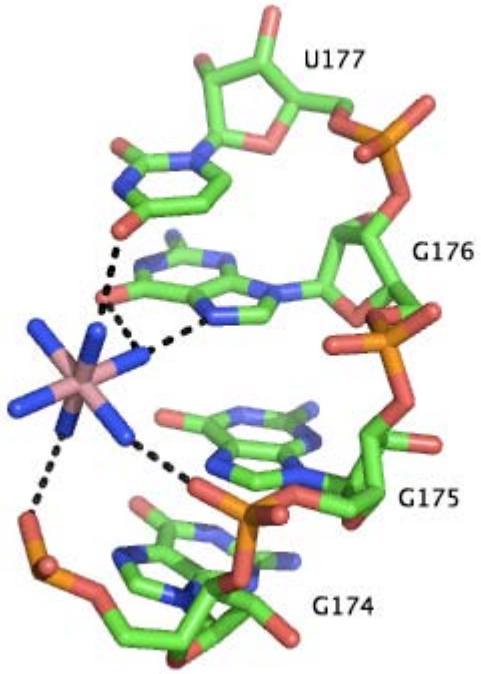
zippers of the CC/AA sequence- two cytosines on the first chain paired to two adenines on the second chain.

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- 5

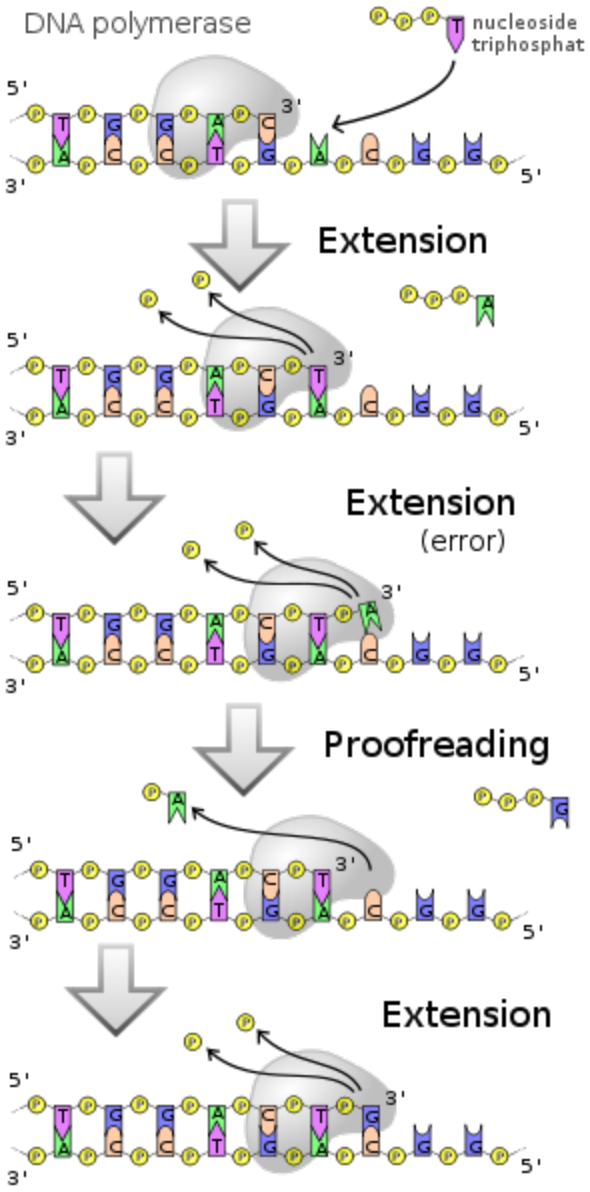
DNA Polymerase

3D structure of the DNA-binding helix-turn-helix motifs in human DNA polymerase beta

A **DNA polymerase** is an enzyme that catalyzes the polymerization of deoxyribonucleotides into a DNA strand. DNA polymerases are best-known for their feedback role in DNA replication, in which the polymerase "reads" an intact DNA strand

as a template and uses it to synthesize the new strand. This process copies a piece of DNA. The newly-polymerized molecule is complementary to the template strand and identical to the template's original partner strand. DNA polymerases use magnesium ions as cofactors. Human DNA polymerases are 900-1000 amino acids long.

Function



DNA polymerase with proofreading ability

DNA polymerase can add free nucleotides to only the 3' end of the newly-forming strand. This results in elongation of the new strand in a 5'-3' direction. No known DNA polymerase is able to begin a new chain (*de novo*). DNA polymerase can add a nucleotide onto only a preexisting 3'-OH group, and, therefore, needs a primer at which it can add

the first nucleotide. Primers consist of RNA and/or DNA bases. In DNA replication, the first two bases are always RNA, and are synthesized by another enzyme called primase. An enzyme known as a helicase is required to unwind DNA from a double-strand structure to a single-strand structure to facilitate replication of each strand consistent with the semiconservative model of DNA replication.

Error correction is a property of some, but not all, DNA polymerases. This process corrects mistakes in newly-synthesized DNA. When an incorrect base pair is recognized, DNA polymerase reverses its direction by one base pair of DNA. The 3'-5' exonuclease activity of the enzyme allows the incorrect base pair to be excised (this activity is known as *proofreading*). Following base excision, the polymerase can re-insert the correct base and replication can continue.

Various DNA polymerases are extensively used in molecular biology experiments.

Variation across species

DNA polymerases have highly-conserved structure, which means that their overall catalytic subunits vary, on a whole, very little from species to species. Conserved structures usually indicate important, irreplaceable functions of the cell, the maintenance of which provides evolutionary advantages.

Some viruses also encode special DNA polymerases, such as Hepatitis B virus DNA polymerase. These may selectively replicate viral DNA through a variety of mechanisms. Retroviruses encode an unusual DNA polymerase called reverse transcriptase, which is an RNA-dependent DNA polymerase (RdDp). It polymerizes DNA from a template of RNA.

DNA polymerase families

Based on sequence homology, DNA polymerases can be further subdivided into seven different families: A, B, C, D, X, Y, and RT.

Family A

Polymerases contain both replicative and repair polymerases. Replicative members from this family include the extensively-studied T7 DNA polymerase, as well as the eukaryotic mitochondrial DNA Polymerase γ. Among the repair polymerases are *Escherichia coli* DNA pol I, *Thermus aquaticus* pol I, and *Bacillus stearothermophilus* pol I. These repair polymerases are involved in excision repair and processing of Okazaki fragments generated during lagging strand synthesis.

Family C

Polymerases are the primary bacterial chromosomal replicative enzymes. DNA Polymerase III alpha subunit from *E. coli* is the catalytic subunit and possesses no known

nuclease activity. A separate subunit, the epsilon subunit, possesses the 3'-5' exonuclease activity used for editing during chromosomal replication. Recent research has classified Family C polymerases as a subcategory of Family X.

Family D

Polymerases are still not very well characterized. All known examples are found in the Euryarchaeota subdomain of Archaea and are thought to be replicative polymerases.

Family X

Contains the well-known eukaryotic polymerase pol β, as well as other eukaryotic polymerases such as pol σ, pol λ, pol μ, and terminal deoxynucleotidyl transferase (TdT). Pol β is required for short-patch base excision repair, a DNA repair pathway that is essential for repairing abasic sites. Pol λ and Pol μ are involved in non-homologous end-joining, a mechanism for rejoining DNA double-strand breaks. TdT is expressed only in lymphoid tissue, and adds "n nucleotides" to double-strand breaks formed during V(D)J recombination to promote immunological diversity. The yeast *Saccharomyces cerevisiae* has only one Pol X polymerase, Pol IV, which is involved in non-homologous end-joining.

Family Y

Y Polymerases differ from others in having a low fidelity on undamaged templates and in their ability to replicate through damaged DNA. Members of this family are hence called translesion synthesis (TLS) polymerases. Depending on the lesion, TLS polymerases can bypass the damage in an error-free or error-prone fashion, the latter resulting in elevated mutagenesis. Xeroderma pigmentosum variant (XPV) patients for instance have mutations in the gene encoding Pol η (eta), which is error-free for UV-lesions. In XPV patients, alternative error-prone polymerases, e.g., Pol ζ (zeta) (polymerase ζ is a B Family polymerase a complex of the catalytic subunit REV3L with Rev7, which associates with Rev1), are thought to be involved in mistakes that result in the cancer predisposition of these patients. Other members in humans are Pol ι (iota), Pol κ (kappa), and Rev1 (terminal deoxycytidyl transferase). In *E. coli*, two TLS polymerases, Pol IV (DINB) and Pol V ($UmuD'_2C$), are known.

Family RT

The reverse transcriptase family contains examples from both retroviruses and eukaryotic polymerases. The eukaryotic polymerases are usually restricted to telomerases. These polymerases use an RNA template to synthesize the DNA strand.

Chapter- 6

Prokaryotic DNA Replication and Eukaryotic DNA Replication

Prokaryotic DNA replication

DNA replication in prokaryotes is exemplified in *E. coli*. It is bi-directional and originates at a single origin of replication (OriC).

Initiation

The initiation of DNA replication is mediated by DnaA, a protein that binds to a region of the origin known as the DnaA box. In *E. coli*, there are 5 DnaA boxes, each of which contains a highly conserved 9 bp consensus sequence 5' - TTATCCACA - 3'. Binding of DnaA to this region causes it to become negatively supercoiled. Following this, a region of OriC upstream of the DnaA boxes (known as DnaB boxes) become melted. There are three of these regions, and each is 13 bp long, and AT-rich (which facilitates melting because less energy is required to break the two hydrogen bonds that form between A and T nucleotides). This region has the consensus sequence 5' - GATCTNTTNTTTT - 3. Melting of the DnaB boxes requires ATP (which is hydrolyzed by DnaA). Following melting, DnaA recruits a hexameric helicase (six DnaB proteins) to opposite ends of the melted DNA. This is where the replication fork will form. Recruitment of helicase requires six DnaC proteins, each of which is attached to one subunit of helicase. Once this complex is formed, an additional five DnaA proteins bind to the original five DnaA proteins to form five DnaA dimers. DnaC is then released, and the prepriming complex is complete. In order for DNA replication to continue, SSB protein is needed to prevent the single strands of DNA from forming any secondary structures and to prevent them from reannealing, and DNA gyrase is needed to relieve the stress (by creating negative supercoils) created by the action of DnaB helicase. The unwinding of DNA by DnaB helicase allows for primase (DnaG) an RNA polymerase to prime each DNA template so that DNA synthesis can begin.

Elongation

Once priming is complete, DNA polymerase III holoenzyme is loaded into the DNA and replication begins. The catalytic mechanism of DNA polymerase III involves the use of two metal ions in the active site, and a region in the active site that can discriminate between deoxyribonucleotides and ribonucleotides. The metal ions are general divalent cations that help the 3' OH initiate a nucleophilic attack onto the alpha phosphate of the deoxyribonucleotide and orient and stabilize the negatively charged triphosphate on the deoxyribonucleotide. Nucleophilic attack by the 3' OH on the alpha phosphate releases pyrophosphate, which is then subsequently hydrolyzed (by inorganic phosphatase) into two phosphates. This hydrolysis drives DNA synthesis to completion.

Furthermore, DNA polymerase III must be able to distinguish between correctly paired bases and incorrectly paired bases. This is accomplished by distinguishing Watson-Crick base pairs through the use of an active site pocket that is complementary in shape to the structure of correctly paired nucleotides. This pocket has a tyrosine residue that is able to form van der Waals interactions with the correctly paired nucleotide. In addition, dsDNA (double stranded DNA) in the active site has a wider and shallower minor groove that permits the formation of hydrogen bonds with the third nitrogen of purine bases and the second oxygen of pyrimidine bases. Finally, the active site makes extensive hydrogen bonds with the DNA backbone. These interactions result in the DNA polymerase III closing around a correctly paired base. If a base is inserted and incorrectly paired, these interactions could not occur due to disruptions in hydrogen bonding and van der Waals interactions.

DNA is read in the 3' → 5' direction, therefore, nucleotides are synthesized (or attached to the template strand) in the 5' → 3' direction. However, one of the parent strands of DNA is 3' → 5' while the other is 5' → 3'. To solve this, replication occurs in opposite directions. Heading towards the replication fork, the leading strand is synthesized in a continuous fashion, only requiring one primer. On the other hand, the lagging strand, heading away from the replication fork, is synthesized in a series of short fragments known as Okazaki fragments, consequently requiring many primers. The RNA primers of Okazaki fragments are subsequently degraded by RNAse H and DNA Polymerase I (exonuclease), and the gap (or nicks) are filled with deoxyribonucleotides and sealed by the enzyme ligase.

Termination

Termination of DNA replication in *E. coli* is completed through the use of termination sequences and the Tus protein. These sequences allow the two replication forks to pass through in only one direction, but not the other.

DNA replication initially produces two catenated or linked circular DNA duplexes, each comprising one parental strand and one newly synthesised strand (by nature of semiconservative replication). This catenation can be visualised as two interlinked rings

which cannot be separated. Topoisomerase IV in *E. coli* unlinks or decatenates the two circular DNA duplexes.

Regulation

Regulation of DNA replication is achieved through several mechanisms. Mechanisms involve the ratio of ATP to ADP, of DnaA to the number of DnaA boxes and the hemimethylation and sequestering of OriC.

The ratio of ATP to ADP indicates that the cell has reached a specific size and is ready to divide. This "signal" occurs because in a rich medium, the cell will grow quickly and will have a lot of excess DNA.

Furthermore, DnaA binds equally well to ATP or ADP, and only the DnaA-ATP complex is able to initiate replication. Thus, in a fast growing cell, there will be more DnaA-ATP than DnaA-ADP. Because the levels of DnaA are strictly regulated, and 5 DnaA-DnaA dimers are needed to initiate replication, the ratio of DnaA to the number of DnaA boxes in the cell is important. After DNA replication is complete, this number is halved, thus DNA replication cannot occur until the levels of DnaA protein increases.

Finally, DNA is sequestered to a membrane-binding protein called SeqA. This protein binds to hemi-methylated GATC DNA sequences. This four bp sequences occurs 11 times in OriC, and newly synthesized DNA only has its parent strand methylated. DAM methyltransferase methylates the newly synthesized strand of DNA only if it is not bound to SeqA. The importance of hemi-methylation is twofold. Firstly, OriC becomes inaccessible to DnaA, and secondly, DnaA binds better to fully methylated DNA than hemi-methylated DNA.

These mechanisms serve to downregulate DNA replication so that it only occurs once per cell cycle, preventing over-replication of DNA.

Eukaryotic DNA replication

DNA replication in eukaryotes is much more complicated than in prokaryotes, although there are many similar aspects. Eukaryotic cells can only initiate DNA replication at a specific point in the cell cycle, the beginning of S phase.

Mechanism

Location in cell cycle

DNA replication in eukaryotes occurs only in the S phase of the cell cycle. However, pre-initiation occurs in the G1 phase. Thus, the separation of pre-initiation and activation ensures that the origin can only fire once per cell cycle.

Due to the sheer size of chromosomes in eukaryotes, eukaryotic chromosomes contain multiple origins of replication. Some origins are well characterized, such as the autonomously replicating sequences (ARS) of yeast while other eukaryotic origins, particularly those in metazoa, can be found in spans of thousands of basepairs. However, the assembly and initiation of replicaton is similar in both the protozoa and metazoa.

Preparation in G1 phase

The first step in DNA replication is the formation of the pre-initiation replication complex (the pre-RC). The formation of this complex occurs in two stages. The first stage requires that there is no CDK activity. This can only occur in early G1. The formation of the pre-RC is known as licensing, but a licensed pre-RC cannot initiate replication in the G1 phase

Current models hold that it begins with the binding of the origin recognition complex (ORC) to the origin. This complex is a hexamer of related proteins and remains bound to the origin, even after DNA replication occurs. Furthermore, ORC is the functional analogue of prokaryotic DnaA. Following the binding of ORC to the origin, Cdc6/Cdc18 and Cdt1 coordinate the loading of the MCM (Mini Chromosome Maintenance) complex to the origin by first binding to ORC and then binding to the MCM complex. The MCM complex is thought to be the major DNA helicase in eukaryotic organisms. Once binding of MCM occurs, a fully licensed pre-RC exists.

Synthesis in S phase

Activation of the complex occurs in S-phase and requires Cdk2-Cyclin E and Ddk. The activation process begins with the addition of Mcm10 to the pre-RC, which displaces Cdt1. Following this, Ddk phosphorylates Mcm3-7, which activates the helicase. It is believed that ORC and Cdc6/18 are phosphorylated by Cdk2-Cyclin E. Ddk and the Cdk complex then recruits another protein called Cdc45, which then recruits all of the DNA

replication proteins to the replication fork. At this stage the origin fires and DNA synthesis begins.

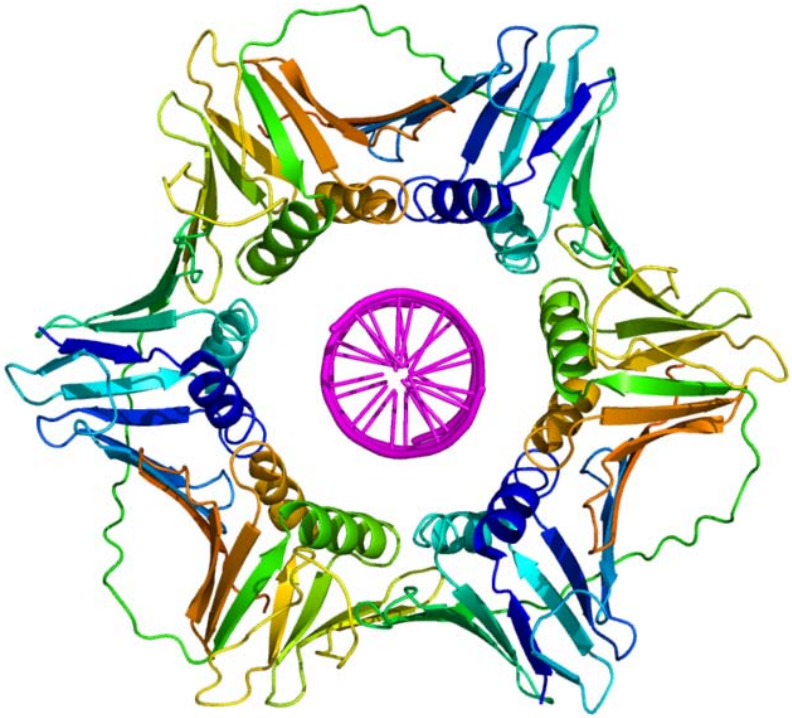
Activation of a new round of replication is prevented through the actions of the cyclin dependent kinases and a protein known as geminin. Geminin binds to Cdt1 and sequesters it. It is a periodic protein that first appears in S-phase and is degraded in late M-phase, possibly through the action of the anaphase promoting complex (APC). In addition, phosphorylation of Cdc6/18 prevent it from binding to the ORC (thus inhibiting loading of the MCM complex) while the phosphorylation of ORC remains unclear. Cells in the G0 stage of the cell cycle are prevented from initiating a round of replication because the Mcm proteins are not expressed.

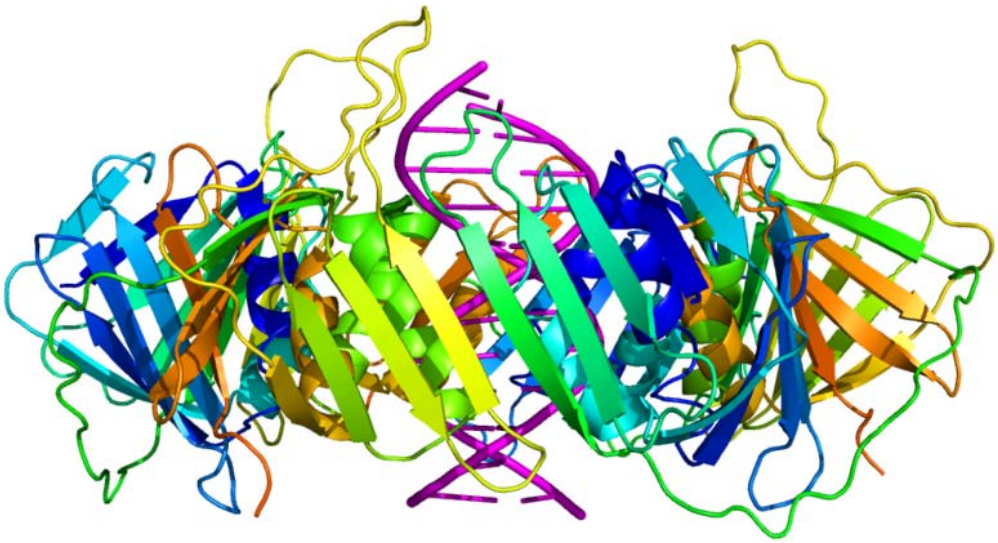
At least three different types of eukaryotic DNA polymerases are involved in the replication of DNA in animal cells (POL α, Pol δ and POL ε).

- **Pol α** forms a complex with a small catalytic (PriS) and a large noncatalytic (PriL) subunit, with the Pri subunits acting as a primase (synthesizing an RNA primer), and then with DNA Pol α elongating that primer with DNA nucleotides. After around 20 nucleotides elongation is taken over by Pol ε (on the leading strand) and δ (on the lagging strand).
- **Pol δ**: Highly processive and has proofreading 3'->5' exonuclease activity. Thought to be the main polymerase involved in lagging strand synthesis, though there is still debate about its role.
- **Pol ε**: Also highly processive and has proofreading 3'->5' exonuclease activity. Highly related to pol δ, and thought to be the main polymerase involved in leading strand synthesis, though there is again still debate about its role.

Chapter- 7

DNA Clamp





Top and side views of a homotrimer of the human PCNA sliding clamp (rainbow colored, N-terminus = blue, C-terminus = red) complexed with double stranded DNA (magenta).

A **DNA clamp**, also known as a **sliding clamp**, is a protein fold that serves as a processivity-promoting factor in DNA replication. As a critical component of the DNA polymerase III holoenzyme, the clamp protein binds DNA polymerase and prevents this enzyme from dissociating from the template DNA strand. The clamp-polymerase protein-protein interactions are stronger and more specific than the direct interactions between the polymerase and the template DNA strand; because the rate-limiting step in the DNA synthesis reaction is the association of the polymerase with the DNA template, the presence of the sliding clamp dramatically increases the number of nucleotides that the polymerase can add to the growing strand per association event. The presence of the DNA clamp can increase the rate of DNA synthesis up to 1,000-fold compared with a nonprocessive polymerase.

Structure

The DNA clamp fold is an α+β protein that assembles into a multimeric structure that completely encircles the DNA double helix as the polymerase adds nucleotides to the growing strand. The DNA clamp assembles on the DNA at the replication fork and "slides" along the DNA with the advancing polymerase, aided by a layer of water molecules in the central pore of the clamp between the DNA and the protein surface. Because of the toroidal shape of the assembled multimer, the clamp cannot dissociate from the template strand without also dissociating into monomers.

The DNA clamp fold is found in bacteria, archaea, eukaryotes and some viruses. In bacteria, the sliding clamp is a homodimer composed of two identical beta subunits of DNA polymerase III and hence is referred to as the beta clamp. In archaea and eukaryotes, it is a trimer composed of three molecules of PCNA. The T4 bacteriophage also uses a sliding clamp, called gp45 that is a trimer similar in structure to PCNA but lacks sequence homology to either PCNA or the bacterial beta clamp.

Kingdom	Sliding clamp protein	Aggregation state	Associated polymerase
Bacteria	beta subunit of pol III	dimer	DNA polymerase III
Archaea	archaeal PCNA	trimer	pol ε
Eukaryote	PCNA	trimer	DNA polymerase delta
Virus	gp43 / gp45	timer	RB69 Pol / T4 Pol

Bacterial

DNA polymerase III subunit beta

Crystallographic structure of the dimeric DNA polymerase beta subunit from *E. coli*.

Identifiers	
Organism	*Escherichia coli*
Symbol	dnaN
Entrez	948218
PDB	1MMI
RefSeq (Prot)	NP_418156
UniProt	P0A988
Other data	
EC number	2.7.7.7
Chromosome	*MG1655: 3.88 - 3.88 Mb*

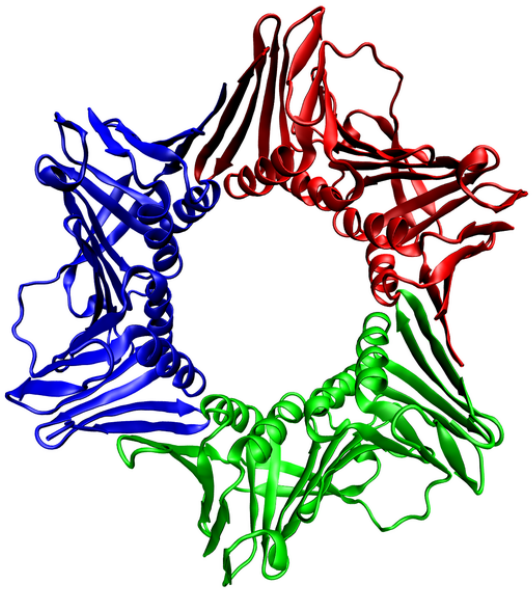
The **beta clamp** is a specific DNA clamp and a subunit of the DNA polymerase III holoenzyme found in bacteria. Two beta subunits are assembled around the DNA by the gamma subunit and ATP hydrolysis; this assembly is called the pre-initiation complex. After assembly around the DNA, the beta subunits' affinity for the gamma subunit is replaced by an affinity for the alpha and epsilon subunits, which together create the complete holoenzyme. DNA polymerase III is the primary enzyme complex involved in prokaryotic DNA replication.

The gamma complex of DNA polymerase III, composed of γδδ'χψ subunits, catalyzes ATP to chaperone two beta subunits to bind to DNA. Once bound to DNA, the beta subunits can freely slide along double stranded DNA. The beta subunits in turn bind the αε polymerase complex. The α subunit possesses DNA polymerase activity and the ε subunit is a 3'-5' exonuclease.

The beta chain of bacterial DNA polymerase III is composed of three topologically non-equivalent domains (N-terminal, central, and C-terminal). Two beta chain molecules are tightly associated to form a closed ring encircling duplex DNA.

Eukaryote

proliferating cell nuclear antigen



The assembled human DNA clamp, a trimer of the human protein PCNA.

Identifiers	
Symbol	PCNA

Entrez	5111
HUGO	8729
OMIM	176740
PDB	1axc
RefSeq	NM_002592
UniProt	P12004
Other data	
EC number	2.7.7.7
Locus	Chr. 20 *pter-p12*

The sliding clamp in eukaryotes is assembled from a specific subunit of DNA polymerase delta called the proliferating cell nuclear antigen (PCNA). The N-terminal and C-terminal domains of PCNA are topologically identical. Three PCNA molecules are tightly associated to form a closed ring encircling duplex DNA.

The sequence of PCNA is well conserved between plants and animals, indicating a strong selective pressure for structure conservation, and suggesting that this type of DNA replication mechanism is conserved throughout eukaryotes. Homologues of PCNA have also been identified in the archaea (*Euryarchaeota* and *Crenarchaeota*) and in Paramecium bursaria Chlorella virus 1 (PBCV-1) and in nuclear polyhedrosis viruses.

Viral

DNA polymerase accessory protein 45

Crystallographic structure of the trimeric gp45 sliding clamp from *bacteriophage T4*.

Identifiers

Organism	*Enterobacteria phage T4*
Symbol	gp45
Entrez	1258821
PDB	1CZD
RefSeq (Prot)	NP_049666
UniProt	P04525
Other data	
EC number	2.7.7.7
Chromosome	*1: 0.03 - 0.03 Mb*

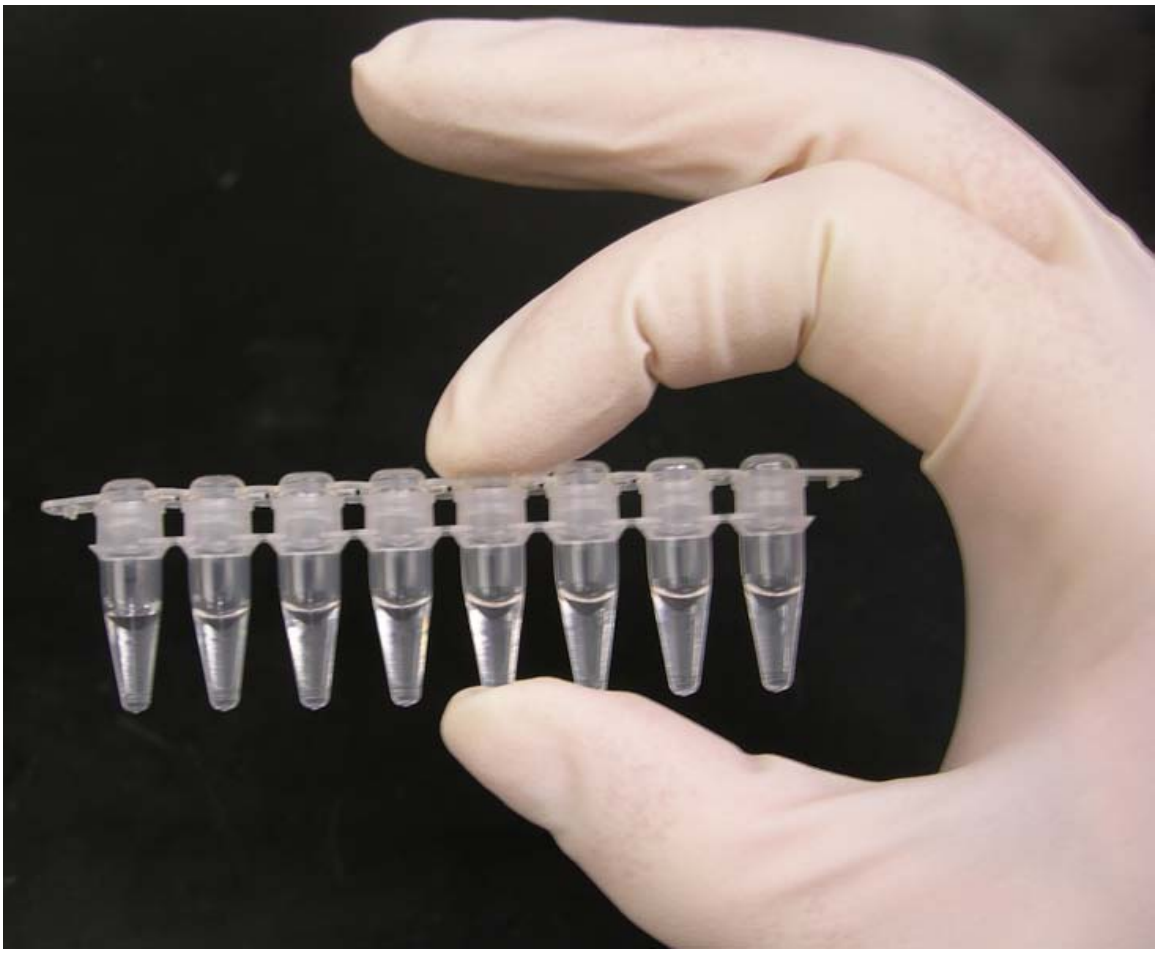
The viral gp45 sliding clamp subunit protein contains two domains. Each domain consists of two alpha helices and two beta sheets - the fold is duplicated and has internal pseudo two-fold symmetry. Three gp45 molecules are tightly associated to form a closed ring encircling duplex DNA.

Assembly

Sliding clamps are loaded onto their associated DNA template strands by specialized proteins known as "sliding clamp loaders", which also disassemble the clamps after replication has completed. The binding sites for these initiator proteins overlap with the binding sites for the DNA polymerase, so the clamp cannot simultaneously associate with a clamp loader and with a polymerase. Thus the clamp will not be actively disassembled while the polymerase remains bound. Although DNA clamps play a less significant role in associating with other DNA-interacting proteins, such as nucleosome assembly factors, Okazaki fragment ligases, and DNA repair proteins, all of these proteins also share a binding site on the DNA clamp that overlaps with the clamp loader site, ensuring that the clamp will not be removed while any enzyme is still working on the DNA. The activity of the clamp loader requires ATP hydrolysis to "close" the clamp around the DNA.

Chapter- 8

Polymerase Chain Reaction



A strip of eight PCR tubes, each containing a 100 µl reaction mixture

The **polymerase chain reaction** (**PCR**) is a scientific technique in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

Developed in 1983 by Kary Mullis, PCR is now a common and often indispensable technique used in medical and biological research labs for a variety of applications. These include DNA cloning for sequencing, DNA-based phylogeny, or functional analysis of genes; the diagnosis of hereditary diseases; the identification of genetic fingerprints (used in forensic sciences and paternity testing); and the detection and diagnosis of infectious diseases. In 1993, Mullis was awarded the Nobel Prize in Chemistry for his work on PCR.

The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA. Primers (short DNA fragments) containing sequences complementary to the target region along with a DNA polymerase (after which the method is named) are key components to enable selective and repeated amplification. As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified. PCR can be extensively modified to perform a wide array of genetic manipulations.

Almost all PCR applications employ a heat-stable DNA polymerase, such as Taq polymerase, an enzyme originally isolated from the bacterium *Thermus aquaticus*. This DNA polymerase enzymatically assembles a new DNA strand from DNA building blocks, the nucleotides, by using single-stranded DNA as a template and DNA oligonucleotides (also called DNA primers), which are required for initiation of DNA synthesis. The vast majority of PCR methods use thermal cycling, i.e., alternately heating and cooling the PCR sample to a defined series of temperature steps. These thermal cycling steps are necessary first to physically separate the two strands in a DNA double helix at a high temperature in a process called DNA melting. At a lower temperature, each strand is then used as the template in DNA synthesis by the DNA polymerase to selectively amplify the target DNA. The selectivity of PCR results from the use of primers that are complementary to the DNA region targeted for amplification under specific thermal cycling conditions.

PCR principles and procedure

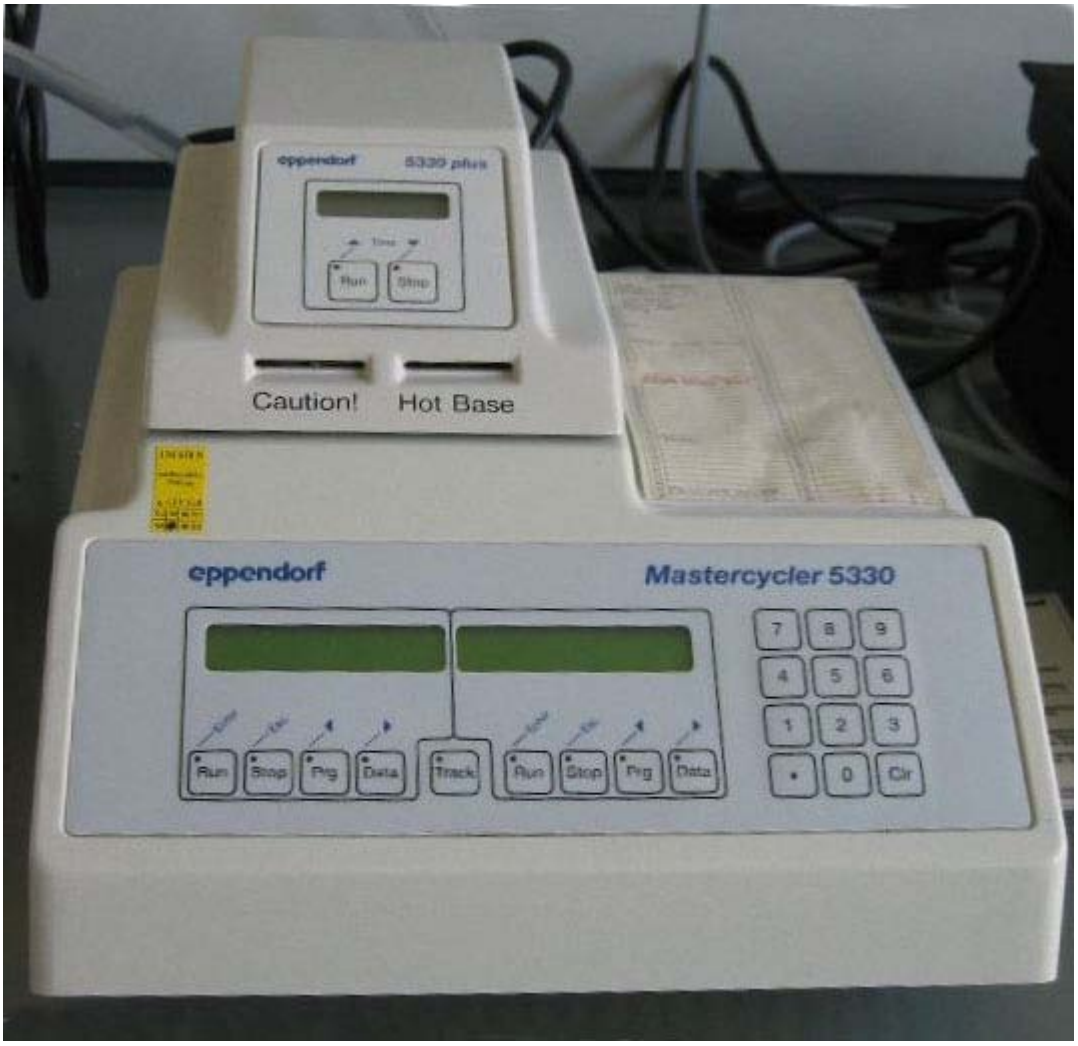


Figure 1a: A thermal cycler for PCR

Figure 1b: An older model three-temperature thermal cycler for PCR

PCR is used to amplify a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of up to ~10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size.

A basic PCR set up requires several components and reagents. These components include:

- *DNA template* that contains the DNA region (target) to be amplified.
- Two *primers* that are complementary to the 3' (three prime) ends of each of the sense and anti-sense strand of the DNA target.

- *Taq polymerase* or another DNA polymerase with a temperature optimum at around 70 °C.
- *Deoxynucleotide triphosphates* (dNTPs), the building blocks from which the DNA polymerases synthesizes a new DNA strand.
- *Buffer solution,* providing a suitable chemical environment for optimum activity and stability of the DNA polymerase.
- *Divalent cations,* magnesium or manganese ions; generally Mg^{2+} is used, but Mn^{2+} can be utilized for PCR-mediated DNA mutagenesis, as higher Mn^{2+} concentration increases the error rate during DNA synthesis
- *Monovalent cation* potassium ions.

The PCR is commonly carried out in a reaction volume of 10–200 µl in small reaction tubes (0.2–0.5 ml volumes) in a thermal cycler. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction (see below). Many modern thermal cyclers make use of the Peltier effect which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled reaction tubes permit favorable thermal conductivity to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermocyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.

Procedure

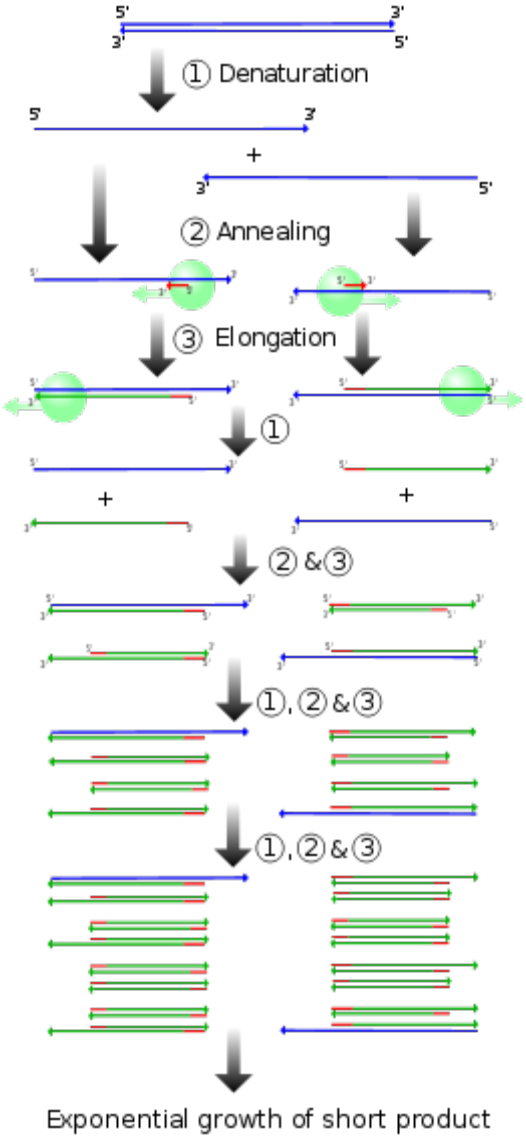


Figure 2: Schematic drawing of the PCR cycle. **(1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C**. Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give shorter DNA products (green lines), which themselves are used as templates as PCR progresses.

Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of 2-3 discrete temperature steps, usually three (Fig. 2). The cycling is often preceded by a single temperature step (called *hold*) at a high temperature (>90°C), and followed by one hold at the end for final product extension or brief storage. The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis,

the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (Tm) of the primers.

- *Initialization step*: This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes. It is only required for DNA polymerases that require heat activation by hot-start PCR.

- *Denaturation step*: This step is the first regular cycling event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules.

- *Annealing step*: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically the annealing temperature is about 3-5 degrees Celsius below the Tm of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis.

- *Extension/elongation step*: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to be amplified. As a rule-of-thumb, at its optimum temperature, the DNA polymerase will polymerize a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment.

- *Final elongation*: This single step is occasionally performed at a temperature of 70–74 °C for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended.

- *Final hold*: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.

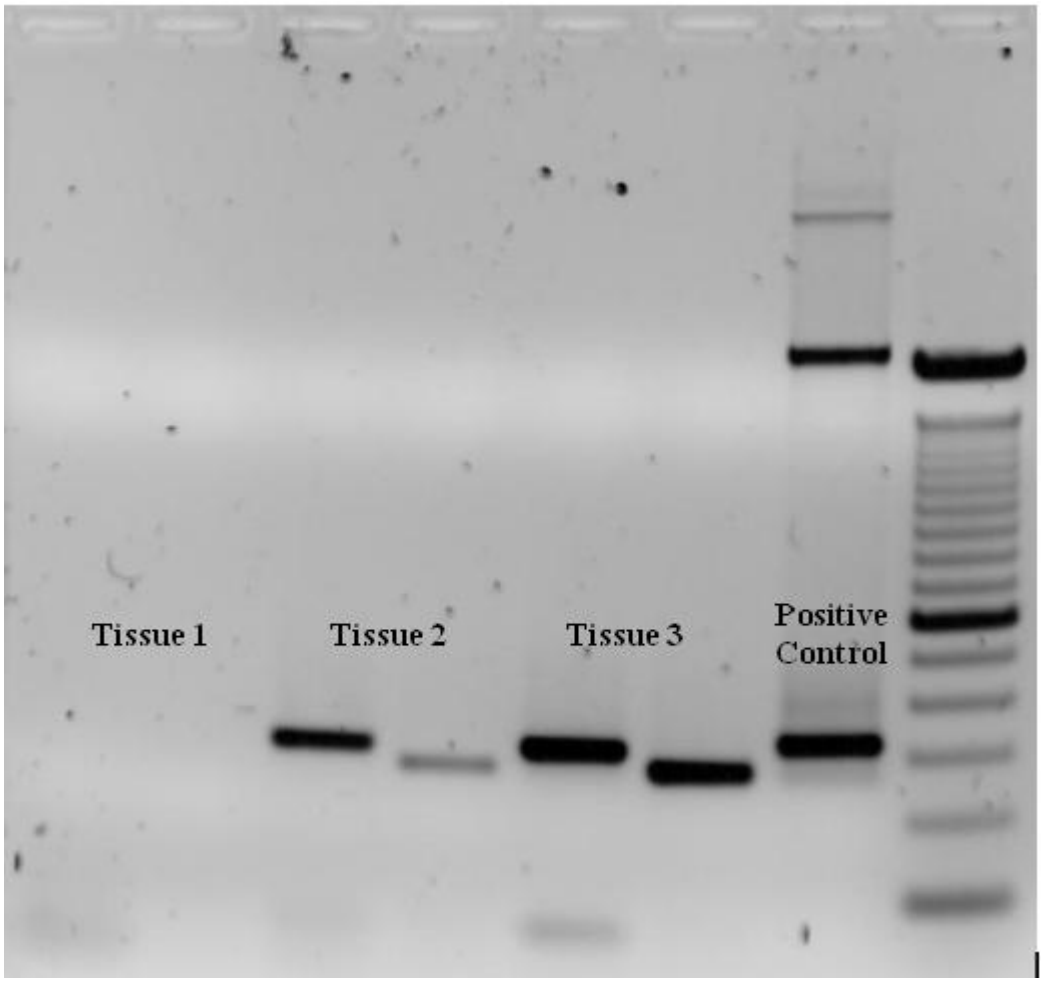


Figure 3: Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. The gel also shows a positive control, and a DNA ladder containing DNA fragments of defined length for sizing the bands in the experimental PCRs.

To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplimer or amplicon), agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products (see Fig. 3).

PCR stages

The PCR process can be divided into three stages:

Exponential amplification: At every cycle, the amount of product is doubled (assuming 100% reaction efficiency). The reaction is very sensitive: only minute quantities of DNA need to be present.

Levelling off stage: The reaction slows as the DNA polymerase loses activity and as consumption of reagents such as dNTPs and primers causes them to become limiting.

Plateau: No more product accumulates due to exhaustion of reagents and enzyme.

PCR optimization

In practice, PCR can fail for various reasons, in part due to its sensitivity to contamination causing amplification of spurious DNA products. Because of this, a number of techniques and procedures have been developed for optimizing PCR conditions. Contamination with extraneous DNA is addressed with lab protocols and procedures that separate pre-PCR mixtures from potential DNA contaminants. This usually involves spatial separation of PCR-setup areas from areas for analysis or purification of PCR products, use of disposable plasticware, and thoroughly cleaning the work surface between reaction setups. Primer-design techniques are important in improving PCR product yield and in avoiding the formation of spurious products, and the usage of alternate buffer components or polymerase enzymes can help with amplification of long or otherwise problematic regions of DNA. Addition of reagents, such as formamide, in buffer systems may increase the specificity and yield of PCR.

Application of PCR

Selective DNA isolation

PCR allows isolation of DNA fragments from genomic DNA by selective amplification of a specific region of DNA. This use of PCR augments many methods, such as generating hybridization probes for Southern or northern hybridization and DNA cloning, which require larger amounts of DNA, representing a specific DNA region. PCR supplies these techniques with high amounts of pure DNA, enabling analysis of DNA samples even from very small amounts of starting material.

Other applications of PCR include DNA sequencing to determine unknown PCR-amplified sequences in which one of the amplification primers may be used in Sanger sequencing, isolation of a DNA sequence to expedite recombinant DNA technologies involving the insertion of a DNA sequence into a plasmid or the genetic material of another organism. Bacterial colonies (E. coli) can be rapidly screened by PCR for correct DNA vector constructs. PCR may also be used for genetic fingerprinting; a forensic technique used to identify a person or organism by comparing experimental DNAs through different PCR-based methods.

Some PCR 'fingerprints' methods have high discriminative power and can be used to identify genetic relationships between individuals, such as parent-child or between

siblings, and are used in paternity testing (Fig. 4). This technique may also be used to determine evolutionary relationships among organisms.

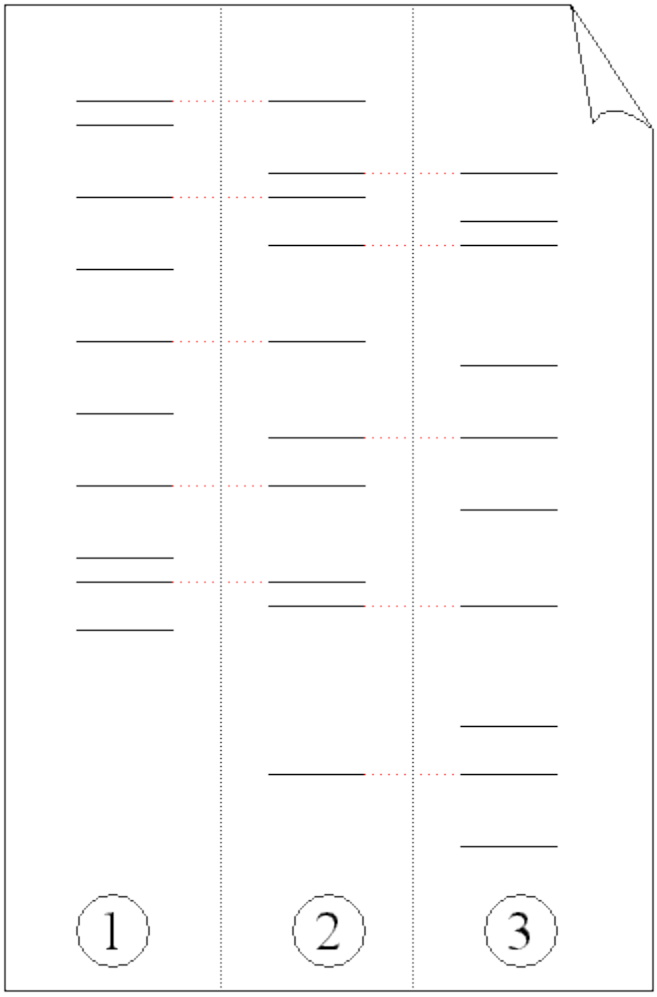


Figure 4: Electrophoresis of PCR-amplified DNA fragments. (1) Father. (2) Child. (3) Mother. The child has inherited some, but not all of the fingerprint of each of its parents, giving it a new, unique fingerprint.

Amplification and quantification of DNA

Because PCR amplifies the regions of DNA that it targets, PCR can be used to analyze extremely small amounts of sample. This is often critical for forensic analysis, when only a trace amount of DNA is available as evidence. PCR may also be used in the analysis of ancient DNA that is tens of thousands of years old. These PCR-based techniques have been successfully used on animals, such as a forty-thousand-year-old mammoth, and also on human DNA, in applications ranging from the analysis of Egyptian mummies to the identification of a Russian tsar.

Quantitative PCR methods allow the estimation of the amount of a given sequence present in a sample—a technique often applied to quantitatively determine levels of gene expression. Real-time PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

PCR in diagnosis of diseases

PCR permits early diagnosis of malignant diseases such as leukemia and lymphomas, which is currently the highest developed in cancer research and is already being used routinely. PCR assays can be performed directly on genomic DNA samples to detect translocation-specific malignant cells at a sensitivity which is at least 10,000 fold higher than other methods.

PCR also permits identification of non-cultivatable or slow-growing microorganisms such as mycobacteria, anaerobic bacteria, or viruses from tissue culture assays and animal models. The basis for PCR diagnostic applications in microbiology is the detection of infectious agents and the discrimination of non-pathogenic from pathogenic strains by virtue of specific genes.

Viral DNA can likewise be detected by PCR. The primers used need to be specific to the targeted sequences in the DNA of a virus, and the PCR can be used for diagnostic analyses or DNA sequencing of the viral genome. The high sensitivity of PCR permits virus detection soon after infection and even before the onset of disease. Such early detection may give physicians a significant lead in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques (see below).

Variations on the basic PCR technique

- *Allele-specific PCR*: a diagnostic or cloning technique which is based on single-nucleotide polymorphisms (SNPs) (single-base differences in DNA). It requires prior knowledge of a DNA sequence, including differences between alleles, and uses primers whose 3' ends encompass the SNP. PCR amplification under stringent conditions is much less efficient in the presence of a mismatch between template and primer, so successful amplification with an SNP-specific primer signals presence of the specific SNP in a sequence.

- *Assembly PCR* or *Polymerase Cycling Assembly (PCA)*: artificial synthesis of long DNA sequences by performing PCR on a pool of long oligonucleotides with short overlapping segments. The oligonucleotides alternate between sense and antisense directions, and the overlapping segments determine the order of the PCR fragments, thereby selectively producing the final long DNA product.

- *Asymmetric PCR*: preferentially amplifies one DNA strand in a double-stranded DNA template. It is used in sequencing and hybridization probing where amplification of only one of the two complementary strands is required. PCR is

carried out as usual, but with a great excess of the primer for the strand targeted for amplification. Because of the slow (arithmetic) amplification later in the reaction after the limiting primer has been used up, extra cycles of PCR are required. A recent modification on this process, known as *L*inear-*A*fter-*T*he-*E*xponential-PCR (LATE-PCR), uses a limiting primer with a higher melting temperature (Tm) than the excess primer to maintain reaction efficiency as the limiting primer concentration decreases mid-reaction.

- *Helicase-dependent amplification*: similar to traditional PCR, but uses a constant temperature rather than cycling through denaturation and annealing/extension cycles. DNA helicase, an enzyme that unwinds DNA, is used in place of thermal denaturation.

- *Hot-start PCR*: a technique that reduces non-specific amplification during the initial set up stages of the PCR. It may be performed manually by heating the reaction components to the melting temperature (e.g., 95°C) before adding the polymerase. Specialized enzyme systems have been developed that inhibit the polymerase's activity at ambient temperature, either by the binding of an antibody or by the presence of covalently bound inhibitors that only dissociate after a high-temperature activation step. Hot-start/cold-finish PCR is achieved with new hybrid polymerases that are inactive at ambient temperature and are instantly activated at elongation temperature.

- *Intersequence-specific PCR* (ISSR): a PCR method for DNA fingerprinting that amplifies regions between simple sequence repeats to produce a unique fingerprint of amplified fragment lengths.

- *Inverse PCR*: is commonly used to identify the flanking sequences around genomic inserts. It involves a series of DNA digestions and self ligation, resulting in known sequences at either end of the unknown sequence.

- *Ligation-mediated PCR*: uses small DNA linkers ligated to the DNA of interest and multiple primers annealing to the DNA linkers; it has been used for DNA sequencing, genome walking, and DNA footprinting.

- *Methylation-specific PCR* (MSP): developed by Stephen Baylin and Jim Herman at the Johns Hopkins School of Medicine, and is used to detect methylation of CpG islands in genomic DNA. DNA is first treated with sodium bisulfite, which converts unmethylated cytosine bases to uracil, which is recognized by PCR primers as thymine. Two PCRs are then carried out on the modified DNA, using primer sets identical except at any CpG islands within the primer sequences. At these points, one primer set recognizes DNA with cytosines to amplify methylated DNA, and one set recognizes DNA with uracil or thymine to amplify unmethylated DNA. MSP using qPCR can also be performed to obtain quantitative rather than qualitative information about methylation.

- *Miniprimer PCR*: uses a thermostable polymerase (S-Tbr) that can extend from short primers ("smalligos") as short as 9 or 10 nucleotides. This method permits PCR targeting to smaller primer binding regions, and is used to amplify conserved DNA sequences, such as the 16S (or eukaryotic 18S) rRNA gene.

- *Multiplex Ligation-dependent Probe Amplification* (*MLPA*): permits multiple targets to be amplified with only a single primer pair, thus avoiding the resolution limitations of multiplex PCR (see below).

- *Multiplex-PCR*: consists of multiple primer sets within a single PCR mixture to produce amplicons of varying sizes that are specific to different DNA sequences. By targeting multiple genes at once, additional information may be gained from a single test run that otherwise would require several times the reagents and more time to perform. Annealing temperatures for each of the primer sets must be optimized to work correctly within a single reaction, and amplicon sizes, i.e., their base pair length, should be different enough to form distinct bands when visualized by gel electrophoresis.

- *Nested PCR*: increases the specificity of DNA amplification, by reducing background due to non-specific amplification of DNA. Two sets of primers are used in two successive PCRs. In the first reaction, one pair of primers is used to generate DNA products, which besides the intended target, may still consist of non-specifically amplified DNA fragments. The product(s) are then used in a second PCR with a set of primers whose binding sites are completely or partially different from and located 3' of each of the primers used in the first reaction. Nested PCR is often more successful in specifically amplifying long DNA fragments than conventional PCR, but it requires more detailed knowledge of the target sequences.

- *Overlap-extension PCR*: a genetic engineering technique allowing the construction of a DNA sequence with an alteration inserted beyond the limit of the longest practical primer length.

- *Quantitative PCR (Q-PCR)*: used to measure the quantity of a PCR product (commonly in real-time). It quantitatively measures starting amounts of DNA, cDNA or RNA. Q-PCR is commonly used to determine whether a DNA sequence is present in a sample and the number of its copies in the sample. *Quantitative real-time PCR* has a very high degree of precision. QRT-PCR methods use fluorescent dyes, such as Sybr Green, EvaGreen or fluorophore-containing DNA probes, such as TaqMan, to measure the amount of amplified product in real time. It is also sometimes abbreviated to RT-PCR (*R*eal *T*ime PCR) or RQ-PCR. QRT-PCR or RTQ-PCR are more appropriate contractions, since RT-PCR commonly refers to reverse transcription PCR, often used in conjunction with Q-PCR.

- *R*everse *T*ranscription PCR (RT-PCR): for amplifying DNA from RNA. Reverse transcriptase reverse transcribes RNA into cDNA, which is then amplified by

PCR. RT-PCR is widely used in expression profiling, to determine the expression of a gene or to identify the sequence of an RNA transcript, including transcription start and termination sites. If the genomic DNA sequence of a gene is known, RT-PCR can be used to map the location of exons and introns in the gene. The 5' end of a gene (corresponding to the transcription start site) is typically identified by RACE-PCR (*Rapid Amplification of cDNA Ends*).

- *Solid Phase PCR*: encompasses multiple meanings, including Polony Amplification (where PCR colonies are derived in a gel matrix, for example), Bridge PCR (primers are covalently linked to a solid-support surface), conventional Solid Phase PCR (where Asymmetric PCR is applied in the presence of solid support bearing primer with sequence matching one of the aqueous primers) and Enhanced Solid Phase PCR (where conventional Solid Phase PCR can be improved by employing high Tm and nested solid support primer with optional application of a thermal 'step' to favour solid support priming).

- *Thermal asymmetric interlaced PCR (TAIL-PCR)*: for isolation of an unknown sequence flanking a known sequence. Within the known sequence, TAIL-PCR uses a nested pair of primers with differing annealing temperatures; a degenerate primer is used to amplify in the other direction from the unknown sequence.

- *Touchdown PCR* (*Step-down PCR*): a variant of PCR that aims to reduce nonspecific background by gradually lowering the annealing temperature as PCR cycling progresses. The annealing temperature at the initial cycles is usually a few degrees (3-5°C) above the T_m of the primers used, while at the later cycles, it is a few degrees (3-5°C) below the primer T_m. The higher temperatures give greater specificity for primer binding, and the lower temperatures permit more efficient amplification from the specific products formed during the initial cycles.

- *PAN-AC*: uses isothermal conditions for amplification, and may be used in living cells.

- *Universal Fast Walking*: for genome walking and genetic fingerprinting using a more specific 'two-sided' PCR than conventional 'one-sided' approaches (using only one gene-specific primer and one general primer - which can lead to artefactual 'noise') by virtue of a mechanism involving lariat structure formation. Streamlined derivatives of UFW are LaNe RAGE (lariat-dependent nested PCR for rapid amplification of genomic DNA ends), 5'RACE LaNe and 3'RACE LaNe.

History

A 1971 paper in the Journal of Molecular Biology by Kleppe and co-workers first described a method using an enzymatic assay to replicate a short DNA template with primers *in vitro*. However, this early manifestation of the basic PCR principle did not receive much attention, and the invention of the polymerase chain reaction in 1983 is generally credited to Kary Mullis.

At the core of the PCR method is the use of a suitable DNA polymerase able to withstand the high temperatures of >90 °C (194 °F) required for separation of the two DNA strands in the DNA double helix after each replication cycle. The DNA polymerases initially employed for in vitro experiments presaging PCR were unable to withstand these high temperatures. So the early procedures for DNA replication were very inefficient, time consuming, and required large amounts of DNA polymerase and continual handling throughout the process.

The discovery in 1976 of Taq polymerase — a DNA polymerase purified from the thermophilic bacterium, *Thermus aquaticus*, which naturally lives in hot (50 to 80 °C (122 to 176 °F)) environments such as hot springs — paved the way for dramatic improvements of the PCR method. The DNA polymerase isolated from *T. aquaticus* is stable at high temperatures remaining active even after DNA denaturation, thus obviating the need to add new DNA polymerase after each cycle. This allowed an automated thermocycler-based process for DNA amplification.

When Mullis developed the PCR in 1983, he was working in Emeryville, California for Cetus Corporation, one of the first biotechnology companies. There, he was responsible for synthesizing short chains of DNA. Mullis has written that he conceived of PCR while cruising along the Pacific Coast Highway one night in his car. He was playing in his mind with a new way of analyzing changes (mutations) in DNA when he realized that he had instead invented a method of amplifying any DNA region through repeated cycles of duplication driven by DNA polymerase. In *Scientific American*, Mullis summarized the procedure: "Beginning with a single molecule of the genetic material DNA, the PCR can generate 100 billion similar molecules in an afternoon. The reaction is easy to execute. It requires no more than a test tube, a few simple reagents, and a source of heat." He was awarded the Nobel Prize in Chemistry in 1993 for his invention, seven years after he and his colleagues at Cetus first put his proposal to practice. However, some controversies have remained about the intellectual and practical contributions of other scientists to Mullis' work, and whether he had been the sole inventor of the PCR principle (see below).

Patent wars

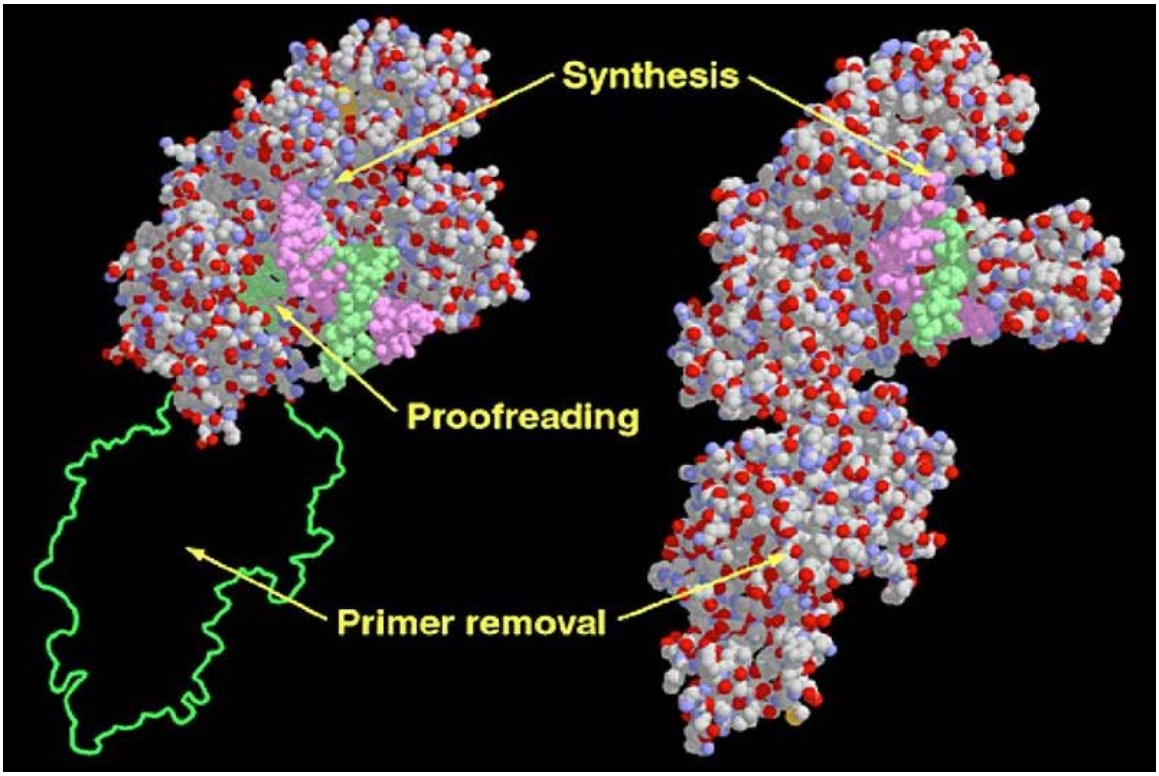
The PCR technique was patented by Kary Mullis and assigned to Cetus Corporation, where Mullis worked when he invented the technique in 1983. The *Taq* polymerase enzyme was also covered by patents. There have been several high-profile lawsuits related to the technique, including an unsuccessful lawsuit brought by DuPont. The pharmaceutical company Hoffmann-La Roche purchased the rights to the patents in 1992 and currently holds those that are still protected.

A related patent battle over the Taq polymerase enzyme is still ongoing in several jurisdictions around the world between Roche and Promega. The legal arguments have extended beyond the lives of the original PCR and Taq polymerase patents, which expired on March 28, 2005.

Chapter- 9

Klenow Fragment and Okazaki Fragment

Klenow fragment



Functional domains in the Klenow Fragment (left) and DNA Polymerase I (PDB).

The **Klenow fragment** is a large protein fragment produced when DNA polymerase I from *E. coli* is enzymatically cleaved by the protease subtilisin. First reported in 1970, it retains the 5'-3' polymerase activity and the 3' → 5' exonuclease activity for removal of precoding nucleotides and proofreading, but loses its 5' → 3' exonuclease activity.

The other smaller fragment formed when DNA polymerase I from *E. coli* is cleaved by subtilisin retains the 5'-3' exonuclease activity but does not have the other two activities exhibited by the Klenow fragment (i.e. 5'-> 3' polymerase activity, and 3'->5' exonuclease activity).

Research

Because the 5' → 3' exonuclease activity of DNA polymerase I from *E. coli* makes it unsuitable for many applications, the Klenow fragment, which lacks this activity, can be very useful in research. The Klenow fragment is extremely useful for research-based tasks such as:

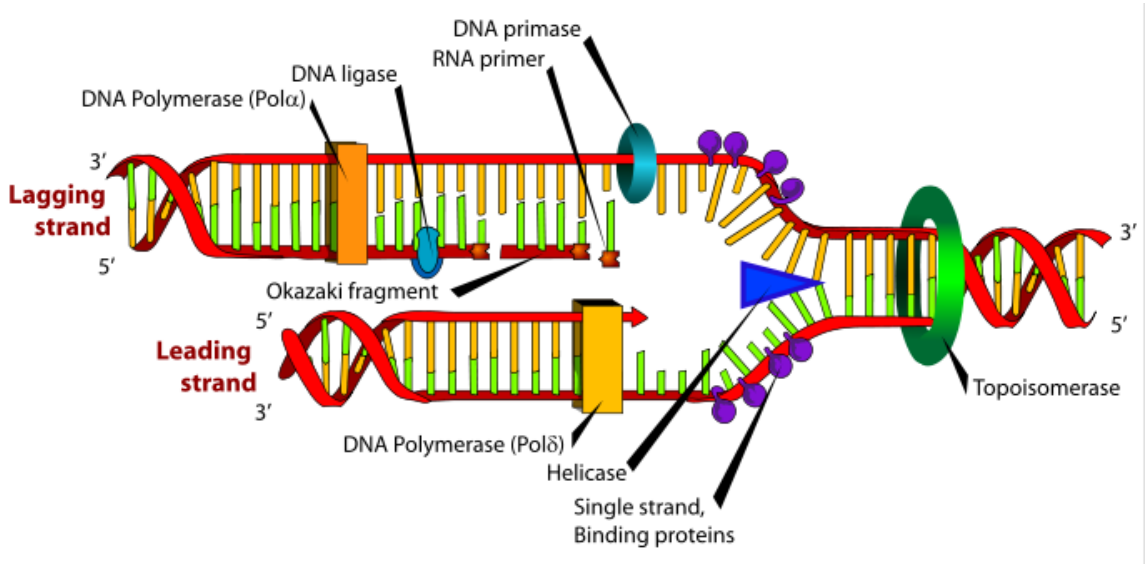
- Synthesis of double-stranded DNA from single-stranded templates
- Filling in (meaning removal of overhangs to create blunt ends) recessed 3' ends of DNA fragments
- Digesting away protruding 3' overhangs
- Preparation of radioactive DNA probes

The Klenow fragment was also the original enzyme used for greatly amplifying segments of DNA in the polymerase chain reaction (PCR) process, before being replaced by thermostable enzymes such as Taq polymerase.

The exo- Klenow fragment

Just as the 5' → 3' exonuclease activity of DNA polymerase I from *E.coli* can be undesirable, the 3' → 5' exonuclease activity of Klenow fragment can also be undesirable for certain applications. This problem can be overcome by introducing mutations in the gene that encodes Klenow. This results in forms of the enzyme being expressed that retain 5' → 3' polymerase activity, but lack any exonuclease activity (5' → 3' or 3' → 5'). This form of the enzyme is called the **exo- Klenow fragment**.

Okazaki fragment



DNA replication

An **Okazaki fragment** is a relatively short fragment of DNA (with no RNA primer at the 5' terminus) created on the lagging strand during DNA replication. The lengths of Okazaki fragments are between 1,000 to 2,000 nucleotides long in *E. coli* and are generally between 100 to 200 nucleotides long in eukaryotes. It was originally discovered in 1968 by Reiji Okazaki, Tsuneko Okazaki, and their colleagues while studying replication of bacteriophage DNA in *Escherichia coli*.

When the lagging strand is being replicated on the original strand, the 5'-3' pattern must be used; thus a small discontinuity occurs and an Okazaki fragment forms. These fragments are processed by the replication machinery to produce a continuous strand of DNA and hence a complete daughter DNA helix.

In dealing with the synthesis of complementary DNA strands the nascent leading strand always reads 3' to 5'. Its antiparallel complement strand, the nascent lagging strand reads from 5' to 3'. Because the original strands of DNA are antiparallel, and only one continuous new strand can be synthesised at the 3' end of the leading strand due to the intrinsic 5'-3' polarity of DNA polymerases, the other strand must grow discontinuously in the opposite direction. Regarding the lagging strand, the result of this strand's discontinuous replication is the production of a series of short sections of DNA called Okazaki fragments.

Each Okazaki fragment is initiated near the replication fork at an RNA primer created by primase, and extended by DNA polymerase III. In eukaryotes, lagging strand synthesis is carried out by the DNA polymerase δ. The primer is later removed by enzymes that have endonucleolytic activity such as Ribonuclease H (RNAse H), flap endonucleases (FENs) and Dna2 helicase/nucleases. In prokaryotes the FEN nuclease is a domain of DNA

polymerase I while in eukaryotes FENs are separate enzymes. The excised RNA bases are replaced with DNA by DNA polymerase I in prokaryotes or DNA polymerase δ in eukaryotes. Adjoining fragments are then linked together by DNA ligase, using phosphodiester bonds, to create a continuous strand of DNA.

Okazaki used a pulse chase type experiment to confirm discontinuous strand replication. He took actively replicating DNA, then added "hot" tritiated nucleotides for a short pulse of about 5 seconds. During the 5 seconds the radioactive nucleotides were incorporated into the growing DNA strands. After the pulse Okazaki chased with "cold" un-labeled nucleotides for varying amounts of time and quickly isolated the DNA. Then the DNA was centrifuged and analyzed for radioactivity. What Okazaki found was that with short chases of about 7 to 15 seconds most of the radioactivity was found in the small fragments higher in the tube after centrifuge. However with longer chases more radioactivity was found in the lower, larger strands. This confirmed that during synthesis first small fragments are formed on the lagging strand, then later these fragments are combined and incorporated into much larger strands. The small fragments found on the lagging strand are called Okazaki fragments.

Chapter- 10

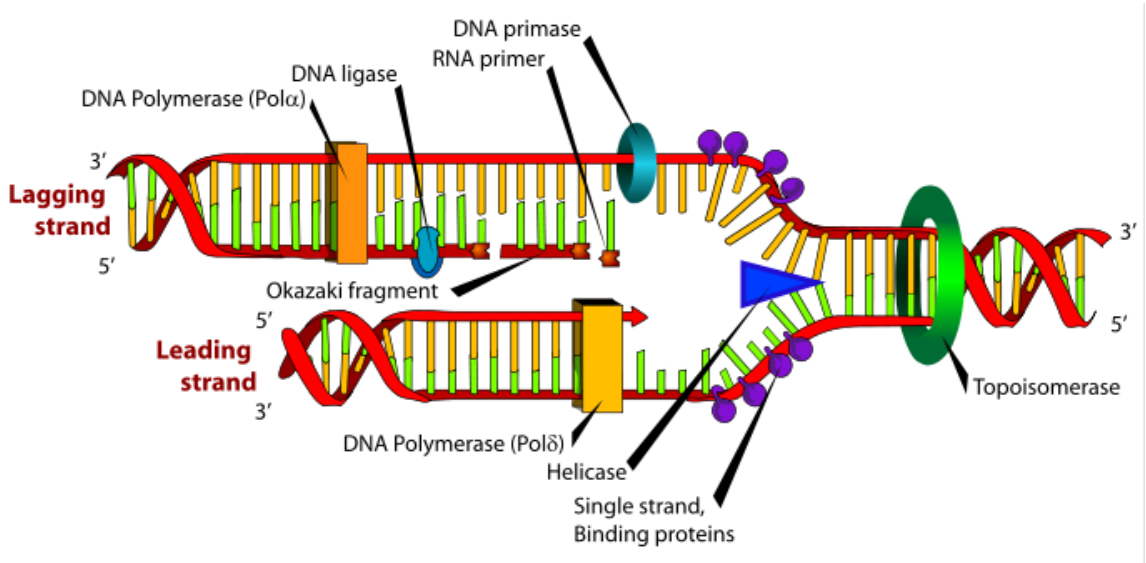
Primer (Molecular Biology)

Primer is a strand of nucleic acid that serves as a starting point for DNA synthesis. They are required because the enzymes that catalyze replication, DNA polymerases, can only add new nucleotides to an existing strand of DNA. The polymerase starts replication at the 3'-end of the primer, and copies the opposite strand.

In most cases of natural DNA replication, the primer for DNA synthesis and replication is a short strand of RNA (which can be made *de novo*).

Many of the laboratory techniques of biochemistry and molecular biology that involve DNA polymerase, such as DNA sequencing and the polymerase chain reaction (PCR), require DNA primers. These primers are usually short, chemically synthesized oligonucleotides, with a length of about twenty bases. They are hybridized to a target DNA, which is then copied by the polymerase.

Mechanism in vivo



The DNA replication fork. RNA primer labeled at top.

The lagging strand is that strand of the DNA double helix that is orientated in a 5' to 3' manner. Therefore, its complement must be synthesized in a 3'→5' manner. Because DNA polymerase III cannot synthesize in the 3'→5' direction, the lagging strand is synthesized in short segments known as Okazaki fragments. Along the lagging strand's template, primase builds RNA primers in short bursts. DNA polymerases are then able to use the free 3'-OH groups on the RNA primers to synthesize DNA in the 5'→3' direction.

The RNA fragments are then removed by DNA polymerase I for prokaryotes or DNA polymerase δ for eukaryotes (different mechanisms are used in eukaryotes and prokaryotes) and new deoxyribonucleotides are added to fill the gaps where the RNA was present. DNA ligase then joins the deoxyribonucleotides together, completing the synthesis of the lagging strand.

Primer removal

In eukaryotic primer removal, DNA polymerase δ extends the Okazaki fragment in 5' to 3' direction, and when it encounters the RNA primer from the previous Okazaki fragment, displacing the 5′ end of the primer into a single-stranded RNA flap, which is removed by nuclease cleavage. Cleavage of the RNA flaps involves either endonuclease 1 (FEN1) cleavage of short flaps, or coating of long flaps by the single-stranded DNA binding protein replication protein A (RPA) and sequential cleavage by Dna2 nuclease and FEN1.

This mechanism is a potential explanation to how HIV virus can transform its genome into double stranded DNA from the RNA-DNA formed after reverse transcription of its RNA. However, the HIV-encoded reverse transcriptase has own ribonuclease activity that degrades the viral RNA during the synthesis of cDNA, as well as DNA-dependent DNA polymerase activity that copies the sense cDNA strand into a *antisense* DNA to form a double-stranded DNA intermediate.

Uses of synthetic primers

DNA sequencing is used to determine the nucleotides in a DNA strand; the chain termination method (dideoxy sequencing or Sanger method) uses a primer as a start marker for the chain reaction.

In PCR, primers are used to determine the DNA fragment to be amplified by the PCR process. The length of primers is usually not more than 30 (usually 18–24) nucleotides, and they need to match the beginning and the end of the DNA fragment to be amplified. They direct replication towards each other – the extension of one primer by polymerase then becomes the template for the other, leading to an exponential increase in the target segment.

It is worth noting that primers are not essentially always necessary for DNA synthesis and can in fact be used by viral polymerases, e.g. influenza, for RNA synthesis.

PCR primer design

Pairs of primers should have similar melting temperatures since annealing in a PCR occurs for both simultaneously. A primer with a T_m significantly higher than the reaction's annealing temperature may mishybridize and extend at an incorrect location along the DNA sequence, while T_m significantly lower than the annealing temperature may fail to anneal and extend at all.

Primer sequences need to be chosen to uniquely select for a region of DNA, avoiding the possibility of mishybridization to a similar sequence nearby. A commonly used method is BLAST search whereby all the possible regions to which a primer may bind can be seen. Both the nucleotide sequence as well as the primer itself can be BLAST searched. The free NCBI tool Primer-BLAST integrates primer design tool and BLAST search into one application, so does commercial software product such as Beacon Designer. Mononucleotide repeats should be avoided, as loop formation can occur and contribute to mishybridization. Primers should not easily anneal with other primers in the mixture (either other copies of same or the reverse direction primer); this phenomenon can lead to the production of 'primer dimer' products contaminating the mixture. Primers should also not anneal strongly to themselves, as internal hairpins and loops could hinder the annealing with the template DNA.

When designing a primer for use in TA cloning, efficiency can be increased by adding AG tails to the 5' and the 3' end.

Degenerate primers

Sometimes *degenerate primers* are used. These are actually mixtures of similar, but not identical primers. They may be convenient if the same gene is to be amplified from different organisms, as the genes themselves are probably similar but not identical. The other use for degenerate primers is when primer design is based on protein sequence. As several different codons can code for one amino acid, it is often difficult to deduce which codon is used in a particular case. Therefore primer sequence corresponding to the amino acid isoleucine might be "ATH", where A stands for adenine, T for thymine, and H for adenine, thymine, or cytosine, according to the genetic code for each codon, using the IUPAC symbols for degenerate bases. Use of degenerate primers can greatly reduce the specificity of the PCR amplification. The problem can be partly solved by using touchdown PCR.

Degenerate primers are widely used and extremely useful in the field of microbial ecology. They allow for the amplification of genes from thus far uncultivated microorganisms or allow the recovery of genes from organisms where genomic information is not available. Usually, degenerate primers are designed by aligning gene sequencing found in GenBank. Differences among sequences are accounted for by using IUPAC degeneracies for individual bases. PCR primers are then synthesized as a mixture of primers corresponding to all permutations.

Chapter- 11

Processivity and Replication Fork

Processivity

In molecular biology, **processivity** is a measure of the average number of nucleotides added by a DNA polymerase enzyme per association/disassociation with the template. DNA polymerases associated with DNA replication tend to be highly processive, while those associated with DNA repair tend to have low processivity. Because the binding of the polymerase to the template is the rate-limiting step in DNA synthesis, the overall rate of DNA replication during S phase of the cell cycle is dependent on the processivity of the DNA polymerases performing the replication. DNA clamp proteins are integral components of the DNA replication machinery and serve to increase the processivity of their associated polymerases. Some polymerases add over 50,000 nucleotides to a growing DNA strand before dissociating from the template strand, giving a replication rate of up to 1,000 nucleotides per second.

DNA binding interactions

Polymerases interact with the phosphate backbone and the minor groove of the DNA, so their interactions do not depend on the specific nucleotide sequence. The binding is largely mediated by electrostatic interactions between the DNA and the "thumb" and "palm" domains of the metaphorically hand-shaped DNA polymerase molecule. When the polymerase advances along the DNA sequence after adding a nucleotide, the interactions with the minor groove dissociate but those with the phosphate backbone remain more stable, allowing rapid re-binding to the minor groove at the next nucleotide.

Interactions with the DNA are also facilitated by DNA clamp proteins, which are multimeric proteins that completely encircle the DNA, with which they associate at replication forks. Their central pore is sufficiently large to admit the DNA strands and some surrounding water molecules, which allows the clamp to slide along the DNA without dissociating from it and without loosening the protein-protein interactions that maintain the toroid shape. When associated with a DNA clamp, DNA polymerase is dramatically more processive; without the clamp most polymerases have a processivity of only about 100 nucleotides. The interactions between the polymerase and the clamp are

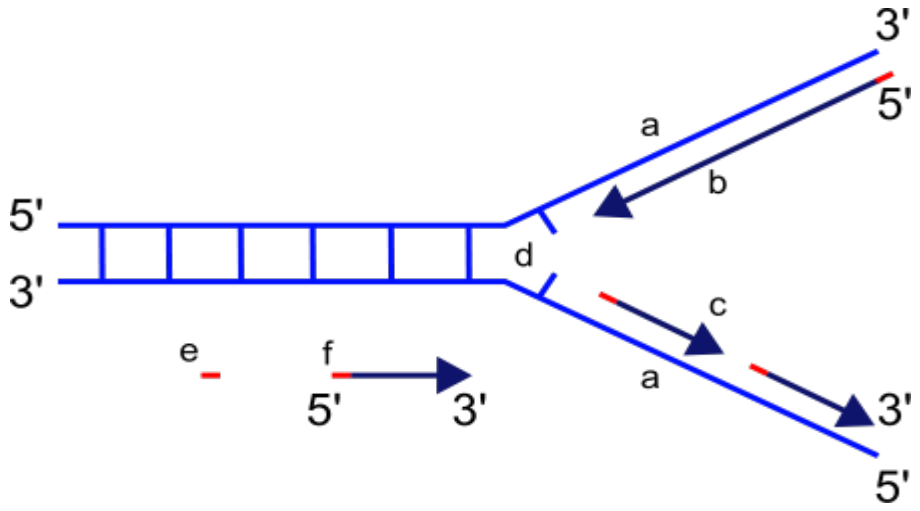
more persistent than those between the polymerase and the DNA. Thus, when the polymerase dissociates from the DNA, it is still bound to the clamp and can rapidly reassociate with the DNA. An example of such a DNA clamp is PCNA (proliferating cell nuclear antigen) found in *S. cervesia*.

Polymerase processivities

Multiple DNA polymerases have specialized roles in the DNA replication process. In *E. coli*, which replicates its entire genome from a single replication fork, the polymerase DNA Pol III is the enzyme primarily responsible for DNA replication and forms a replication complex with extremely high processivity. The related DNA Pol I has exonuclease activity and serves to degrade the RNA primers used to initiate DNA synthesis. Pol I then synthesizes the short DNA fragments that were formerly hybridized to the RNA fragment. Thus Pol I is much less processive than Pol III because its primary function in DNA replication is to create many short DNA regions rather than a few very long regions.

In eukaryotes, which have a much higher diversity of DNA polymerases, the low-processivity initiating enzyme is called Pol α, and the high-processivity extension enzymes are Pol δ and Pol ε. Both prokaryotes and eukaryotes must "trade" bound polymerases to make the transition from initiation to elongation. This process is called polymerase switching.

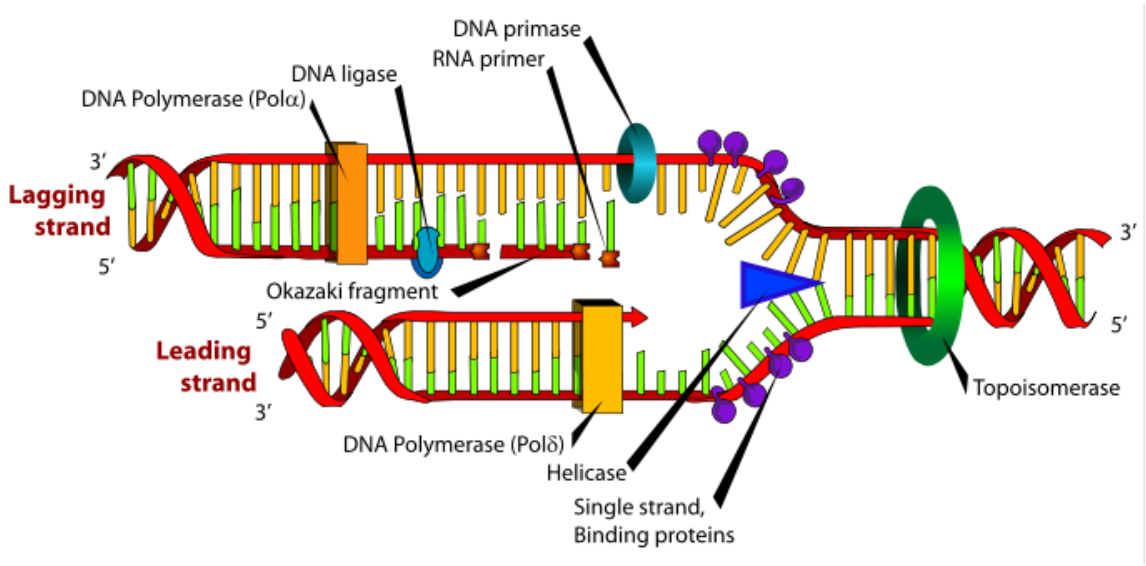
Replication fork



Scheme of the replication fork.
a: template, b: leading strand, c: lagging strand, d: replication fork, e: primer, f: Okazaki fragments

The **replication fork** is a structure that forms within the nucleus during DNA replication. It is created by helicases, which break the hydrogen bonds holding the two DNA strands together. The resulting structure has two branching "prongs", each one made up of a single strand of DNA. These two strands serve as the template for the leading and lagging strands which will be created as DNA polymerase matches complementary nucleotides to the templates. The templates may be properly referred to as the leading strand template and the lagging strand template.

Replication



DNA replication

When replicating, the original DNA splits in two, forming two "prongs" which resemble a fork (hence the name "replication fork"). DNA has a ladder-like structure; imagine a ladder broken in half vertically, along the steps. Each half of the ladder now requires a new half to match it. Because DNA polymerase can only synthesize a new DNA strand in a 5' to 3' manner, the process of replication goes differently for the two strands comprising the DNA double helix.

Leading strand

The **leading strand template** is the template strand of the DNA double helix that is oriented in a 3' to 5' manner. All DNA synthesis occurs 5'-3'. The original DNA strand must be read 3'-5' to produce a 5'-3' nascent strand.

The leading strand is formed along the leading strand template as a polymerase "reads" the template DNA and continuously adds nucleotides to the 3' end of the elongating strand. This polymerase is DNA polymerase III (DNA Pol III) in prokaryotes and presumably Pol ε in eukaryotes.

Lagging strand

The **lagging strand template** is the coding strand of the DNA double helix that is oriented in a 5' to 3' manner. The newly made lagging strand still is synthesized 5'-3'. However, since the DNA is oriented in a manner that does not allow continual synthesis, only small sections can be read at a time. An RNA primer is placed on the DNA strand 3' to the origin of replication. Just as before, DNA Polymerase reads 3'-5' on the original DNA to produce a 5'-3' nascent strand. Polymerase reaches the origin of replication and stops replication until a new RNA primer is placed 3' to the last RNA primer. These fragments of DNA produced on the lagging strand are called Okazaki fragments. The orientation of the original DNA on the lagging strand prevents continual synthesis. As a result, replication of the lagging strand is more complicated than of the leading strand.

On the lagging strand template, primase "reads" the DNA and adds RNA to it in short, separated segments. In eukaryotes, primase is intrinsic to Pol α. DNA polymerase III or Pol δ lengthens the primed segments, forming Okazaki fragments. Primer removal in eukaryotes is also performed by Pol δ. In prokaryotes, DNA polymerase I "reads" the fragments, removes the RNA using its flap endonuclease domain, and replaces the RNA nucleotides with DNA nucleotides (this is necessary because RNA and DNA use slightly different kinds of nucleotides). DNA ligase joins the fragments together.

Chapter- 12

Replication Timing

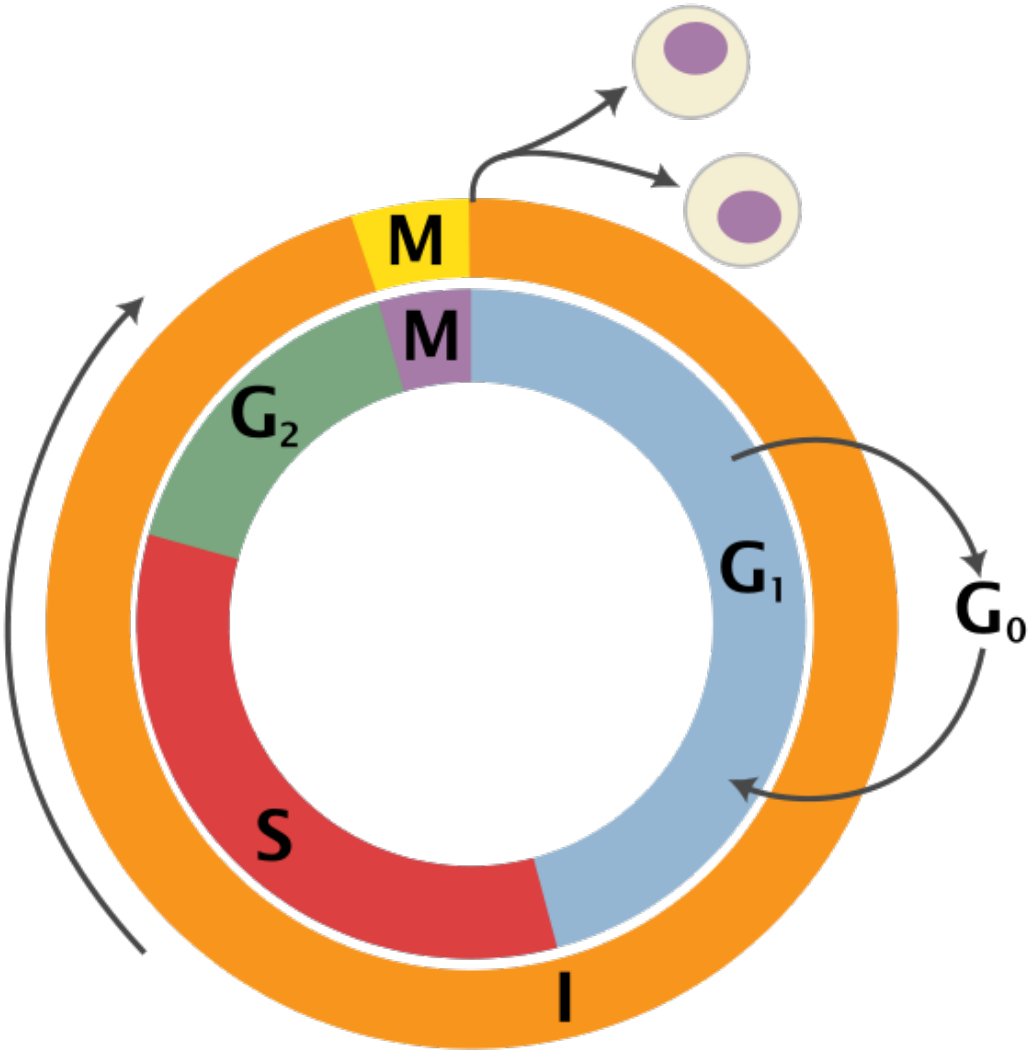


Figure 1: Schematic of the cell cycle. outer ring: I = Interphase, M = Mitosis; inner ring: M = Mitosis, G_1 = Gap 1, G_2 = Gap 2, S = Synthesis; not in ring: G_0 = Gap 0/Resting.

Replication Timing refers to the order in which segments of DNA along the length of a chromosome are duplicated.

DNA Replication

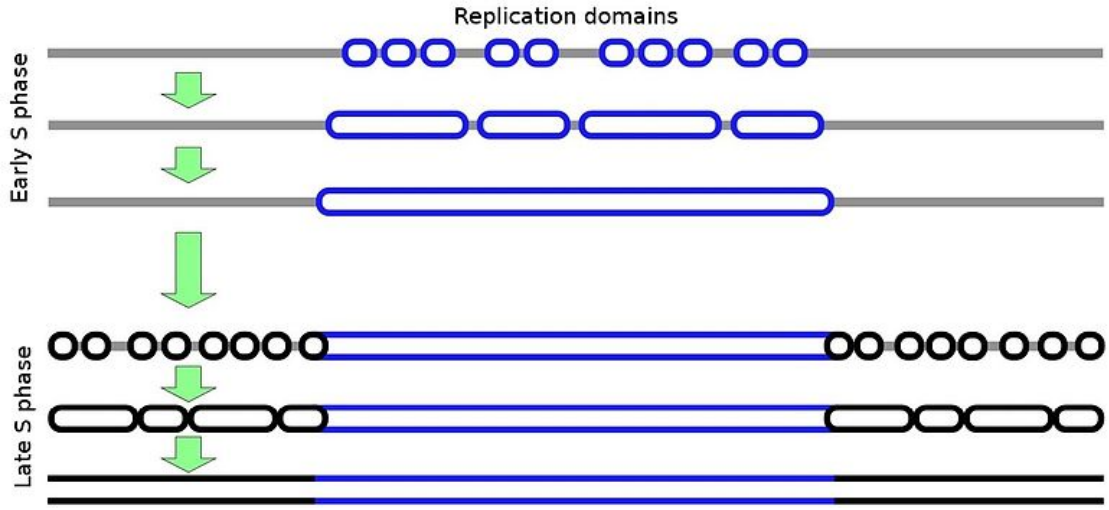


Figure 2: Replication proceeds via the nearly synchronous firing of clusters of replication origins that replicate segments of chromosomal DNA ("Replication domains") at defined time periods during S phase.

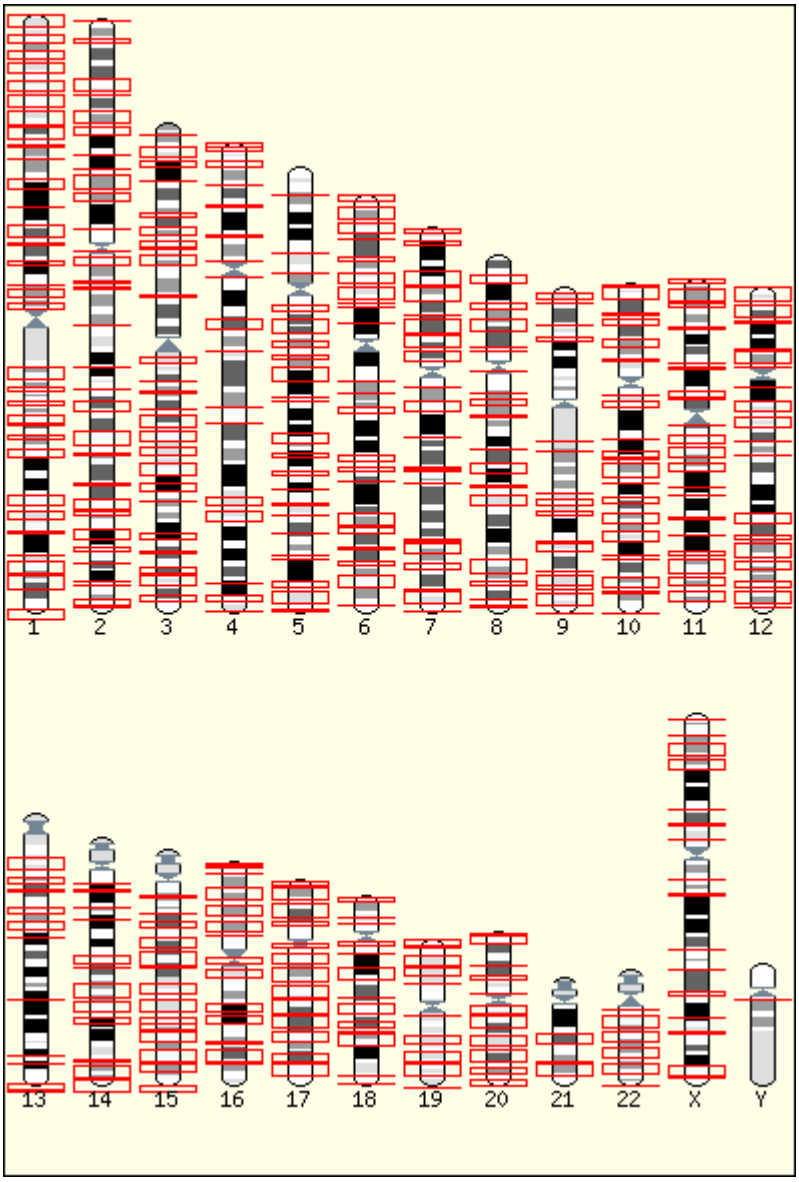


Figure 3: Sequence of replication.

In eukaryotic cells (cells that package their DNA within a nucleus), chromosomes consist of very long linear double-stranded DNA molecules. During the S-phase of each cell cycle (**Figure 1**), all of the DNA in a cell is duplicated in order to provide one copy to each of the daughter cells after the next cell division. The process of duplicating DNA is called DNA replication, and it takes place by first unwinding the duplex DNA molecule, starting at many locations called DNA replication origins, followed by an unzipping process that unwinds the DNA as it is being copied. However, replication does not start at all the different origins at once. Rather, there is a defined temporal order in which these origins fire. Frequently a few adjacent origins open up to duplicate a segment of a chromosome, followed some time later by another group of origins opening up in an adjacent segment. Replication does not necessarily start at exactly the same origin sites every time, but the segments appear to replicate in the same temporal sequence regardless

of exactly where within each segment replication starts. **Figure 2** shows a cartoon of how this is generally envisioned to occur, while **Figure 3** shows when different segments replicate in one type of human cell.

Replication Timing Profiles

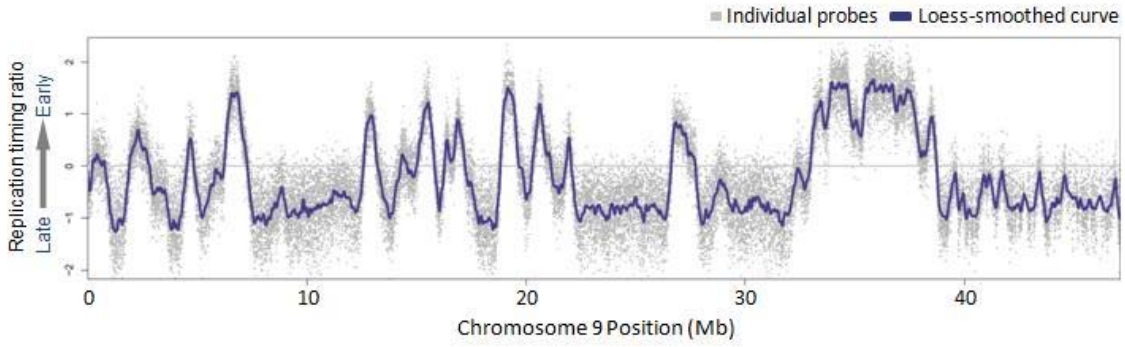


Figure 4: Example of a " Replication Timing Profile". Each data point (grey dot) represents a different DNA sequence position along the length of chromosome 9, indicated on the x axis. The y-axis indicates the time at which each DNA sequence position replicates during S phase, with more positive values being replicated earlier. A smooth curve is drawn through the data points to more easily visualize the domains of different replication timing.

The temporal order of replication of all the segments in the genome, called its replication-timing program, can now be easily measured in two different ways. One way literally and simply measures the amount of the different DNA sequences along the length of the chromosome per cell. Sequences that duplicate first, long before cell division, will be more abundant in each cell than the sequences that replicate last just prior to cell division. The other way is to label newly synthesized DNA with chemically tagged nucleotides that become incorporated into the strands as they are synthesized, and then catch cells at different times during the duplication process and purify the DNA synthesized at each of these times using the chemical tag. In either case, we can measure the amount of the different DNA sequences along the length of the chromosome either directly using a machine that reads how much of each sequence is present or indirectly using a process called microarray hybridization. In any case, the temporal order of replication along the length of each chromosome can be plotted in graphical form to produce a "replication timing profile". **Figure 4** shows an example of such a profile across 50,000,000 base pairs of human Chromosome 9 in a human embryonic stem cell line.

Replication Timing and Chromosome Structure

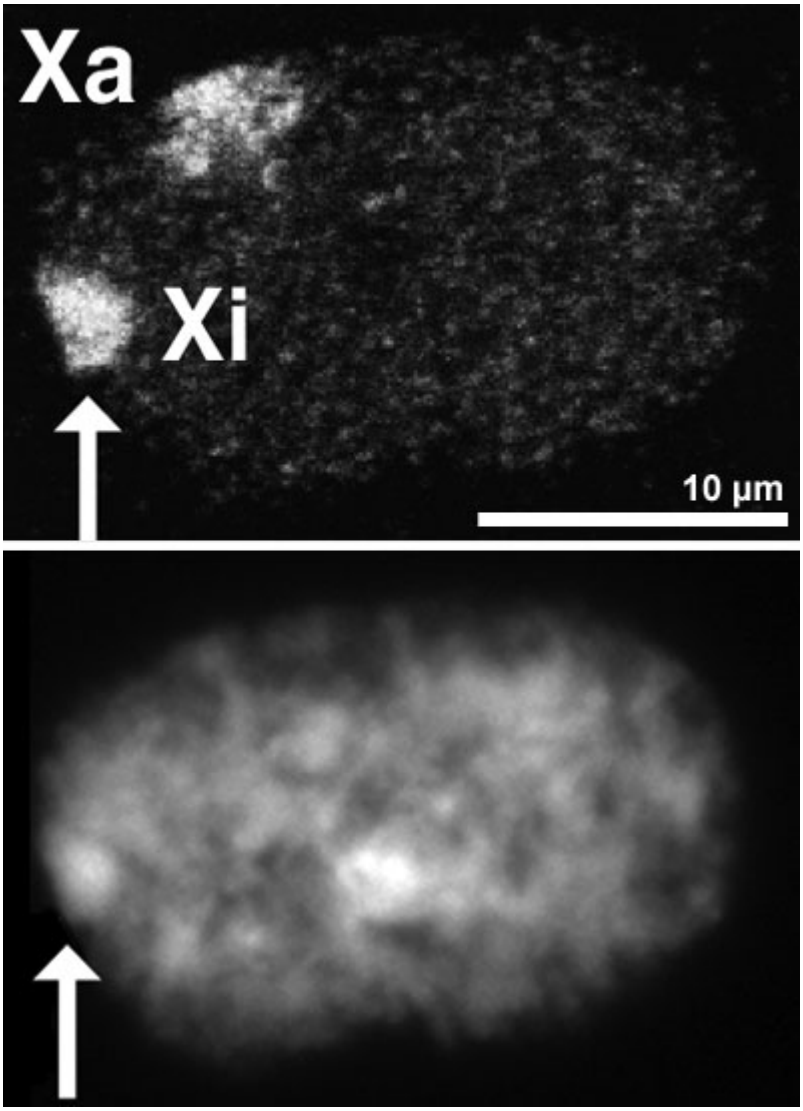


Figure 5. Nucleus of a female amniotic fluid cell. Top: Both X-chromosome territories are detected by FISH. Shown is a single optical section made with a confocal microscope. Bottom: Same nucleus stained with DAPI and recorded with a CCD camera. The Barr body is indicated by the arrow, it identifies the inactive X (Xi).

At present, very little is known about either the mechanisms orchestrating the timing program or its biological significance. However, it is an intriguing cellular mechanism with links to many poorly understood features of the folding of chromosomes inside the cell nucleus. All eukaryotes have a timing program, and this program is similar in related species. This indicates that it is either important itself, or something important influences the program; in other words, it either represents or reflects something that Mother Nature wants to retain. It is unlikely that replicating DNA in a specific temporal order is necessary simply for the basic purpose of duplicating a DNA molecule. More than likely, it is related to some other chromosomal property or function. Replication timing is

correlated with the expression of genes such that the genetic information being utilized in a cell is generally replicated earlier than the information that is not being used. We also know that the replication-timing program changes during development, along with changes in the expression of genes.

For many decades now, it has been known that replication timing is correlated with the structure of chromosomes. For example, female mammals have two X chromosomes. One of these is genetically active, while the other is inactivated early in development. In 1960, J. H. Taylor showed that the active and inactive X chromosomes replicate in a different pattern, with the active X replicating earlier than the inactive X, whereas all the other pairs of chromosomes replicate in the same temporal pattern. It was also noticed by Mary Lyon that the inactive X took on a condensed structure in the nucleus called the Barr body (**Figure 5**) at the same time during development as the genetic inactivation of the chromosome.

Altogether, this may not come as too much of a surprise, since the packaging of DNA with proteins and RNA into chromatin takes place immediately after the DNA is synthesized. Therefore, replication timing dictates the time of assembly of chromatin. Less intuitive is the relationship between replication timing and the 3 dimensional positioning of chromatin in the nucleus. It is now well-accepted that chromatin is not randomly organized in the cell nucleus, but the positions of each chromosome domain relative to its neighboring domains is characteristic of different cell types and after this geography is established in each newly formed cell, the chromosome domains do not move appreciably until the next cell division. Intriguingly, in all multi-cellular organisms where it has been measured, early replication takes place in the interior of the nucleus and the chromatin around the periphery is replicated later. This compartmentalization seems to be quite profound, as recently developed methods to measure the points where different parts of chromosomes touch each other are almost perfectly aligned to when they replicate. In other words, regions that are replicated early vs. late are packaged in such a way as to be spatially segregated in the nucleus, with the intervening DNA containing the regions devoid of origin activity. One possibility is that these different compartments within the nucleus, established and maintained without the aid of membranes or physical barriers, set thresholds for the initiation of replication so that the more accessible regions are the first to replicate.

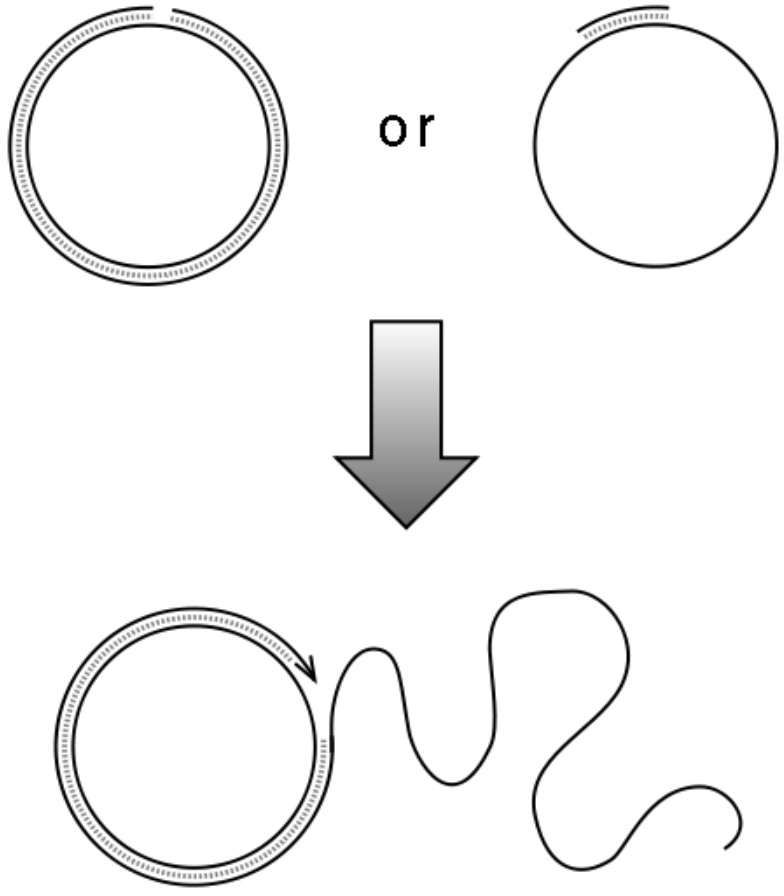
Replication Timing and Disease

Another intriguing aspect of replication timing is that the temporal order of replication is disrupted in most cancers and in many diseases. We do not yet understand the mechanisms behind this link, but it suggests that further research may reveal replication-timing changes as useful biomarkers for such diseases. The fact that it can now be measured with relative ease indicates that we will soon have a wealth of information about where and when large changes in chromosome folding occur during development and in different diseases.

Chapter- 13

Rolling Circle Replication and Semiconservative Replication

Rolling circle replication



Rolling circle replication produces multiple copies of a single circular template.

Rolling circle replication describes a process of unidirectional nucleic acid replication that can rapidly synthesize multiple copies of circular molecules of DNA or RNA, such as plasmids, the genomes of bacteriophages, and the circular RNA genome of viroids. Some eukaryotic viruses also replicate their DNA via a rolling circle mechanism.

Circular DNA replication

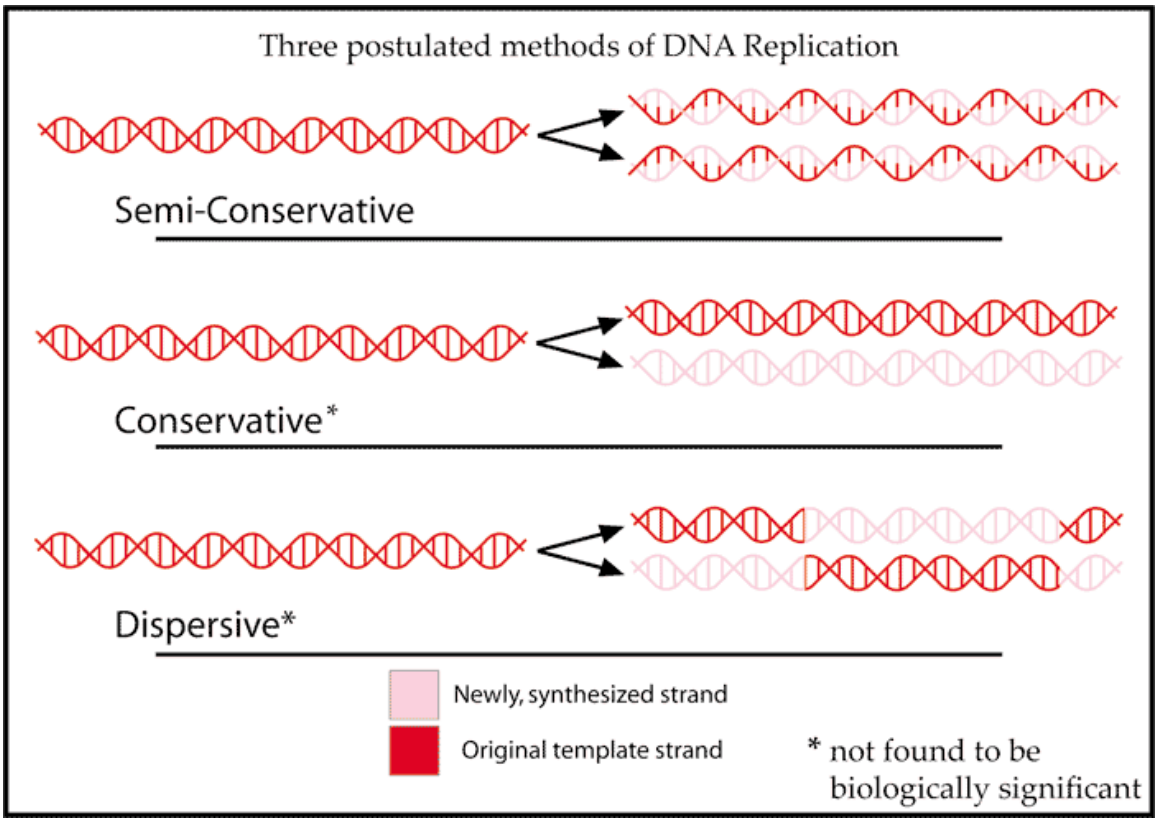
Rolling circle DNA replication is initiated by an initiator protein encoded by the plasmid or bacteriophage DNA, which nicks one strand of the double-stranded, circular DNA molecule at a site called the double-strand origin, or DSO. The initiator protein remains bound to the 5' phosphate end of the nicked strand, and the free 3' hydroxyl end is released to serve as a primer for DNA synthesis by DNA polymerase III. Using the unnicked strand as a template, replication proceeds around the circular DNA molecule, displacing the nicked strand as single-stranded DNA. Displacement of the nicked strand is carried out by a host-encoded helicase called PcrA (the abbreviation standing for plasmid copy reduced) in the presence of the plasmid replication initiation protein.

Continued DNA synthesis can produce multiple single-stranded linear copies of the original DNA in a continuous head-to-tail series called a concatemer. These linear copies can be converted to double-stranded circular molecules through the following process:

First, the initiator protein makes another nick to terminate synthesis of the first (leading) strand. RNA polymerase and DNA polymerase III then replicate the single-stranded origin (SSO) DNA to make another double-stranded circle. DNA polymerase I removes the primer, replacing it with DNA, and DNA ligase joins the ends to make another molecule of double-stranded circular DNA.

Rolling circle replication has found wide uses in academic research and biotechnology, and has been successfully used for amplification of DNA from very small amounts of starting material.

Semiconservative replication



A summary of the three postulated methods of DNA synthesis

Semiconservative replication describes the method by which DNA is replicated in all known cells. This method of replication is one of three proposed models of DNA replication:

- Semiconservative replication would produce two copies that each contained one of the original strands and one new strand.
- Conservative replication would leave the two original template DNA strands together in a double helix and would produce a copy composed of two new strands containing all of the new DNA base pairs.
- Dispersive replication would produce two copies of the DNA, both containing distinct regions of DNA composed of either both original strands or both new strands.

The deciphering of the structure of DNA by Watson and Crick in 1953 suggested that each strand of the double helix would serve as a template for synthesis of a new strand. However, there was no way of knowing how the newly synthesized strands might combine with the template strands to form two double helical DNA molecules. The semiconservative model seemed most reasonable since it would allow each daughter strand to remain associated with its template strand. The semiconservative model was

confirmed by the *Meselson-Stahl experiment* and other even more revealing experiments that allowed for autoradiographic visualization of the distribution of old and new strands within replicated chromosomes.

Testing the semi-conservative theory

Biophysical evidence

The semi-conservative theory can be confirmed by making use of the fact that DNA is made up of nitrogen bases. Nitrogen has an isotope N15 (N14 is the most common isotope) called heavy nitrogen. The experiment that confirms the predictions of the semi-conservative theory makes use of this isotope and runs as follows:

1. Bacterial (E coli) DNA is placed in a media containing heavy nitrogen(N15), which binds to the DNA, making it identifiable.
2. This DNA is then placed in a media with the presence of N14 and left to replicate only once. The new bases will contain nitrogen 14 while the originals will contain N15
3. The DNA is placed in test tubes containing caesium chloride (heavy compound) and centrifuged at 40,000 revolutions per minute.
4. The caesium chloride molecules sink to the bottom of the test tubes creating a density gradient. The DNA molecules will position at their corresponding level of density (taking into account that N15 is more dense than N14)
5. These test tubes are observed under ultraviolet rays. DNA appears as a fine layer in the test tubes at different heights according to their density.

According to the semi-conservative theory, after one replication of DNA, we should obtain 2 hybrid (part N14 part N15) molecules from each original strand of DNA. This would appear as a single line in the test tube. This result would be the same for the dispersive theory. On the other hand, according to the conservative theory, we should obtain one original DNA strand and a completely new one i.e. two fine lines in the test tube placed separately one from the other. Up to this point, either the semi-conservative or the dispersive theories could be truthful, as experimental evidence confirmed that only one line appeared after one replication. In order to conclude between those two, DNA had to be left to replicate again, still in a media containing N14.

In the dispersive theory, after 2 divisions we should obtain a single line, but further up in the test tube, as the DNA molecules become less dense as N14 becomes more abundant in the molecule According to the semi-conservative theory, 2 hybrid molecules and 2 fully N14 molecules should be produced, so two fine lines at different heights in the test tubes should be observed. Experimental evidence confirmed that two lines were observed therefore offering compelling evidence for the semi-conservative theory.

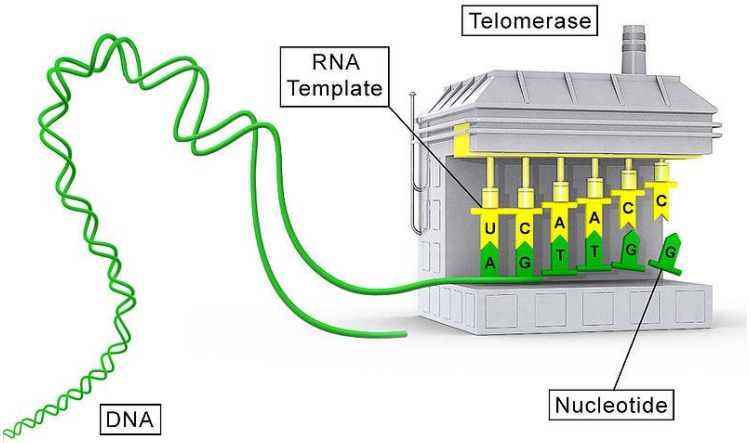
Genetic evidence

An independent 'genetic' evidence for the semi-conservative theory was provided more recently by high throughput genomic sequencing of individual mutagenized bacteria. E. coli were treated with Ethyl methanesulfonate (EMS), known to induce G:C → A:T transitions due to generation of abnormal base O-6-ethylguanine, which is further misrecognized during DNA replication and paired with T instead of C. The sequenced DNA from individual colonies of EMS-mutagenized bacteria exhibited long stretches of solely G → A or C → T transitions, which in some cases were spanning entire bacterial genome. The elementary explanation of this observation is based on semi-conservative mechanism: one should expect the segregation between daughter strands into different cells after replication, which leads to each descendant cell having exclusively G → A or C → T conversions.

Chapter- 14

Telomerase

RNA-directed DNA polymerase



A conceptual diagram showing the protein component of telomerase (TERT) in grey and the RNA component (TR) in yellow

Identifiers	
EC number	2.7.7.49
CAS number	9068-38-6
Databases	
IntEnz	IntEnz view
BRENDA	BRENDA entry
ExPASy	NiceZyme view

KEGG	KEGG entry
MetaCyc	metabolic pathway
PRIAM	profile
PDB	structures
Gene Ontology	AmiGO / EGO

Telomerase is an enzyme that adds DNA sequence repeats ("TTAGGG" in all vertebrates) to the 3' end of DNA strands in the telomere regions, which are found at the ends of eukaryotic chromosomes. This region of repeated nucleotide repeats called telomeres contain non-coding DNA material and prevents constant loss of important DNA from chromosome ends. As a result, every time the chromosome is copied only 100-200 nucleotides are lost, which causes no damage to the organism's DNA. Telomerase is a reverse transcriptase that carries its own RNA molecule, which is used as a template when it elongates telomeres, which are shortened after each replication cycle.

The existence of a compensatory shortening of telomere (telomerase) mechanism was first predicted by Soviet biologist Alexey Olovnikov in 1973, who also suggested the telomere hypothesis of aging and the telomere's connections to cancer. Telomerase was discovered by Carol W. Greider and Elizabeth Blackburn in 1984 in the ciliate *Tetrahymena*. Together with Jack W. Szostak, Greider and Blackburn were awarded the 2009 Nobel Prize in Physiology or Medicine for their discovery.

Structure

The protein composition of human telomerase was identified in 2007 by Scott Cohen and his team at the Children's Medical Research Institute in Australia. It consists of two molecules each of human telomerase reverse transcriptase (TERT), telomerase RNA (TR or TERC), and dyskerin (DKC1). The genes of telomerase subunits, which are TERT, TERC, DKC1, and TEP1 *etc*, are located on the different chromosomes in human genome. Human TERT gene (hTERT) is translated into a protein of 1132 amino acids. TERT proteins from many eukaryotes have been sequenced. TERT polypeptide folds with TERC, a non-coding RNA (451 nucleotides long in human). TERT has a 'mitten' structure that allows it to wrap around the chromosome to add single-stranded telomere repeats.

TERT is a reverse transcriptase, which is a class of enzyme that creates single-stranded DNA using single-stranded RNA as a template. Enzymes of this class (not TERT specifically, but the ones isolated from viruses) are utilized by scientists in the molecular biological process of reverse transcriptase PCR (RT-PCR), which allows the creation of several DNA copies of a target sequence using RNA as a template. As stated above, TERT carries its own template around, TERC.

The high-resolution protein structure of the *Tribolium castaneum* catalytic subunit of telomerase TERT was decoded in 2008 by Emmanuel Skordalakes and his team at The

Wistar Institute in Philadelphia. The structure revealed that the protein consists of four conserved domains (RNA-Binding Domain (TRBD), fingers, palm and thumb), organized into a ring configuration that shares common features with retroviral reverse transcriptases, viral RNA polymerases and bacteriophage B-family DNA polymerases.

Function

By using TERC, TERT can add a six-nucleotide repeating sequence, 5'-TTAGGG (in all vertebrates, the sequence differs in other organisms) to the 3' strand of chromosomes. These TTAGGG repeats (with their various protein binding partners) are called telomeres. The template region of TERC is 3'-CAAUCCCAAUC-5'. This way, telomerase can bind the first few nucleotides of the template to the last telomere sequence on the chromosome, add a new telomere repeat (5'-GGTTAG-3') sequence, let go, realign the new 3'-end of telomere to the template, and repeat the process.

Clinical implications

Aging

The enzyme telomerase allows for replacement of short bits of DNA known as telomeres, which are otherwise shortened when a cell divides via mitosis.

In normal circumstances, without the presence of telomerase, if a cell divides recursively, at some point all the progeny will reach their Hayflick limit. With the presence of telomerase, each dividing cell can replace the lost bit of DNA, and any single cell can then divide unbounded. While this unbounded growth property has excited many researchers, caution is warranted in exploiting this property, as exactly this same unbounded growth is a crucial step in enabling cancerous growth.

Embryonic stem cells express telomerase, which allows them to divide repeatedly and form the individual. In adults, telomerase is highly expressed in cells that need to divide regularly (e.g., in the immune system), whereas most somatic cells express it only at very low levels in a cell-cycle-dependent manner.

A variety of premature aging syndromes are associated with short telomeres. These include Werner syndrome, Ataxia telangiectasia, Ataxia-telangiectasia like disorder, Bloom syndrome, Fanconi anemia and Nijmegen breakage syndrome. The genes that have been mutated in these diseases all have roles in the repair of DNA damage, and their precise roles in maintaining telomere length are an active area of investigation. While it is currently unknown to what extent telomere erosion contributes to the normal aging process, maintenance of DNA in general and telomeric DNA, to be specific, have emerged as major players. Dr. Michael Fossel has suggested in an interview that telomerase therapies may be used not only to combat cancer but also to actually get around human aging and extend lifespan significantly. He believes human trials of telomerase-based therapies for extending lifespan will occur within the next 10 years.

This timeline is significant because it coincides with the retirement of Baby Boomers in the United States and Europe.

Some experiments have raised questions on whether telomerase can be used as an anti-aging therapy, namely, the fact that mice with elevated levels of telomerase have higher cancer incidence and hence do not live longer. In addition, although certain premature aging syndromes have been associated with telomere shortening, mice without active telomerase do not appear to suffer from premature aging. Telomerase also favors tumorogenesis, leading to questions about its potential as an anti-aging therapy. On the other hand, one study showed that activating telomerase in cancer-resistant mice by overexpressing its catalytic subunit extended lifespan. The potential remains for telomerase activators to contribute to the development of cancer.

Exposure of T lymphocytes from HIV-infected human donors to a small molecule telomerase activator (TAT2) retards telomere shortening, increases proliferative potential, and, importantly, enhances cytokine/chemokine production and antiviral activity.

A study which focused on Ashkenazi Jews, found that those that live the longest inherit a hyperactive version of telomerase that rebuilds telomeres.

In mice engineered to block the gene that produces telomerase unless they are given a certain drug aged at a much faster rate and died at about six months, instead of the average mouse lifespan of about three years. Administering the drug at 6 months turned on telomerase production and caused their organs to be "rejuvenated," restored fertility, and normalized their ability to detect or process odors. The finding raises hope for treatment of conditions such as progeria and other accelerated aging disorders, as well as possible organ regeneration therapies, such as repair of liver damage due to hepatitis or alcoholism.

A study published in the journal *Nature* in January 2011 found that Telomerase reactivation reversed tissue degeneration in older telomerase-deficient mice.

Cancer

When cells are approaching the Hayflick limit in cell cultures, the time to senescence can be extended by the inactivation of the tumor suppressor proteins - TP53 and Retinoblastoma protein (pRb). Cells that have been so-altered will eventually undergo an event termed a "crisis" when the majority of the cells in the culture die. Sometimes, a cell does not stop dividing once it reaches crisis. In a typical situation, the telomeres are lost, and the integrity of the chromosomes declines with every subsequent cell division. Exposed chromosome ends are interpreted as double-stranded breaks (DSB) in DNA; such damage is usually repaired by reattaching (religating) the broken ends together. When the cell does this due to telomere-shortening, the ends of different chromosomes can be attached together. This temporarily solves the problem of lacking telomeres; but, during anaphase of cell division, the fused chromosomes are randomly ripped apart, causing many mutations and chromosomal abnormalities. As this process continues, the

cell's genome becomes unstable. Eventually, either sufficient damage will be done to the cell's chromosomes such that cell dies (via programmed cell death, apoptosis), or an additional mutation that activates telomerase will take place.

With the activation of telomerase, some types of cells and their offspring become immortal, that is, their chromosomes will not become unstable no matter how many cell divisions they undergo (they bypass the Hayflick limit), thus avoiding cell death as long as the conditions for their duplication are met. Many cancer cells are considered 'immortal' because telomerase activity allows them to divide virtually forever, which is why they can form tumors. A good example of cancer cells' immortality is HeLa cells, which have been used in laboratories as a model cell line since 1951.

While this method of modeling human cancer in cell culture is effective and has been used for many years by scientists, it is also very imprecise. The exact changes that allow for the formation of the tumorigenic clones in the above-described experiment are not clear. Scientists have subsequently been able to address this question by the serial introduction of several mutations present in a variety of human cancers. This has led to the elucidation of several combinations of mutations that are sufficient for the formation of tumorigenic cells, in a variety of cell types. While the combination varies depending on the cell type, a common theme is that the following alterations are required: activation of TERT, loss of p53 pathway function, loss of pRb pathway function, activation of the Ras or myc proto-oncogenes, and aberration of the PP2A protein phosphatase. That is to say, the cell has an activated telomerase, eliminating the process of death by chromosome instability or loss, absence of apoptosis-induction pathways, and continued activation of mitosis.

This model of cancer in cell culture accurately describes the role of telomerase in actual human tumors. Telomerase activation has been observed in ~90% of all human tumors, suggesting that the immortality conferred by telomerase plays a key role in cancer development. Of the tumors that have not activated TERT, most have found a separate pathway to maintain telomere length termed ALT (Alternative Lengthening of Telomeres). The exact mechanism behind telomere maintenance in the ALT pathway has not been elucidated, but likely involves multiple recombination events at the telomere.

Additional roles in cancer, heart disease, and quality of life

Additional roles for telomerase per work by Elizabeth Blackburn *et al.*, include the upregulation of 70 genes known or suspected in cancers' growth and spread through the body, and the activation of glycolysis, which enables cancer cells to rapidly use sugar to facilitate their programmed growth rate (roughly the growth rate of a fetus).

E. V. Gostjeva *et al.* (MIT) recently imaged colon cancer stem cells and compared them to fetal colon stem cells trying to make a new colon; they were the same.

Elizabeth Blackburn *et al.* UCSF has shown work that reveals that mothers caring for their very sick children have shorter telomeres when they report that their emotional

stress is at the greatest point. She also found telomerase active at the site of blockages in coronary artery tissue. This could be why heart attacks can come on so suddenly: Telomerase is driving the growth of the blockage.

In 2009, it was shown that the amount of telomerase activity significantly increased due to psychological stress. Across the sample of patients telomerase activity increased by 18% one hour after the end of the stress. Telomerase activity was examined in peripheral blood mononuclear cells.

According to a 2007 study, there is no correlation between socio-economic status and telomere length.

Blackburn and the two other co-discoverers of telomerase won the Lasker Award (2006), and the Nobel Prize (2009) for the discovery of telomerase and subsequent work on telomerase. Blackburn also won the 2006 Gruber Genetics Prize for same.

Role in other human diseases

Mutations in TERT have been implicated in predisposing patients to aplastic anemia, a disorder in which the bone marrow fails to produce blood cells, in 2005.

Cri du chat Syndrome (CdCS) is a complex disorder involving the loss of the distal portion of the short arm of chromosome 5. TERT is located in the deleted region, and loss of one copy of TERT has been suggested as a cause or contributing factor of this disease.

Dyskeratosis congenita (DC) is a disease of the bone marrow that can be caused by some mutations in the telomerase subunits. In the DC cases, about 35% cases are X-linked-recessive on the DKC1 locus and 5% cases are autosomal dominant on the TERT and TERC loci.

Patients with DC have severe bone marrow failure manifesting as abnormal skin pigmentation, leucoplakia (a white thickening of the oral mucosa), and nail dystrophy, as well as a variety of other symptoms. Individuals with either TERC or DKC1 mutations have shorter telomeres and defective telomerase activity *in vitro* than other individuals of the same age.

There has also been one family in which autosomal dominant DC has been linked to a heterozygous mutation in TERT. These patients also exhibited an increased rate of telomere-shortening, and genetic anticipation (i.e., the DC phenotype worsened with each generation).

Telomerase as a potential drug target

Cancer is a very difficult disease to fight because the immune system has trouble recognizing it, and cancer cells are immortal; they will always continue dividing. Because telomerase is necessary for the immortality of so many cancer types, it is thought to be a

potential drug target. If a drug can be used to turn off telomerase in cancer cells, the above process of telomere-shortening will resume—telomere length will be lost as the cells continue to divide, mutations will occur, and cell stability will decrease. Experimental drug and vaccine therapies targeting active telomerase have been tested in mouse models, and some have now entered early clinical trials. Geron Corporation is currently conducting four human clinical trials involving telomerase inhibition and telomerase vaccination. Merck, as a licensee of Geron, has recent approval of an IND for one vaccine type. The vaccine platform is being tested (and now jointly with Merck) using three different approaches. One vaccine is adenovirus/plasmid based (Merck IND). The second is an autologous dendritic cell based vaccine (GRNVAC1), formerly called TVAX when tested in Phase I clinical trials in Prostate Cancer, and it showed significant PSA doubling times as well as T-cell response. Geron's embryonic stem cell derived dendritic cell vaccine targeting telomerase is the third approach and is currently at the pre-clinical stage. These vaccine methods attempt to teach the human immune system to attack cancer cells expressing telomerase. Geron's telomerase inhibitor drug (GRN163L) attempts to stop cancer cell proliferation by inhibiting telomerase and it is in three separate early stage human clinical trials. Indeed, telomerase inhibition in many types of cancer cells grown in culture has led to the massive death of the cell population. However, a variety of caveats, including the presence of the ALT pathway, complicate such therapies. Some have reported ALT methods of telomere maintenance and storage of DNA in cancer stem cells, however Geron claims to have killed cancer stem cells with their telomerase inhibitor GRN163L at Johns Hopkins. GRN163L binds directly to the RNA template of telomerase. Even a mutation of the RNA template of telomerase would render the telomerase unable to extend telomeres, and therefore not be able to grant replicative immortality to cancer, not allow glycolysis to be inititated, and not upregulate Blackburn's 70 cancer genes. Since Blackburn has shown that most of the harmful cancer-related effects of telomerase are dependent on an intact RNA template, it seems a very worthwhile target for drug development. If indeed some cancer stem cells use an alternative method of telomere maintenance, it should be noted that they are still killed when the RNA template of telomerase is blocked. According to Blackburn's opinion at most of her lectures, it is a big mistake to think that telomerase is involved with only extending telomeres. Stopping glycolysis in cancer stem cells and preventing the upregulation of 70 bad genes is probably what is killing cancer stem cells if they are using alternative methods.