Subject: Bioinformatics

B.Sc 1st Year (Semester-I)

Course Title: Foundations of Bioinformatics-I

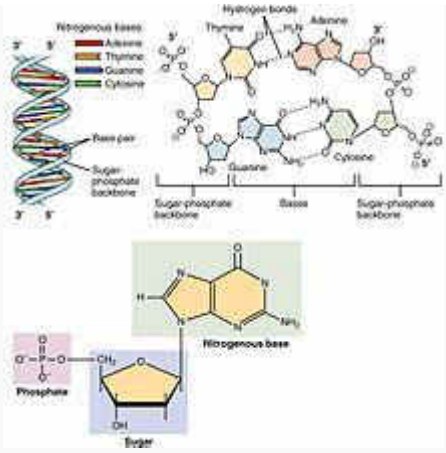
Unit-II

Nucleotides are organic molecules that serve as the monomers, or subunits, of nucleic acids like DNA (deoxyribonucleic acid) and RNA(ribonucleic acid). The building blocks of nucleic acids, nucleotides are composed of a nitrogenous base, a five-carbon sugar (ribose or deoxyribose), and at least one phosphate group. Thus a nucleoside plus a phosphate group yields a nucleotide.

Nucleotides also function to carry packets of energy within the cell in the form of the nucleoside triphosphates (ATP, GTP, CTP and UTP), playing a central role in metabolism. In addition, nucleotides participate in cell signaling (cGMP and cAMP), and are incorporated into important cofactors of enzymatic reactions (e.g. coenzyme A, FAD, FMN, NAD, and NADP).

In experimental biochemistry, nucleotides can be radiolabeled with radionuclides to yield radionucleotides.

Structure:



The structure of nucleotide monomers

A nucleotide is made of a nucleobase (also termed a nitrogenous base), a five-carbon sugar (either ribose or 2-deoxyribose, depending on if it is RNA or DNA), and one or, depending on the definition, more than one phosphate groups. Authoritative chemistry sources such as the ACS Style Guideand IUPAC Gold Book clearly state that the term nucleotide refers only to a molecule containing one phosphate. However, common usage in molecular biology textbooks often extends this definition to include molecules with two or three phosphate groups. Thus, the term "nucleotide" generally refers to a nucleoside monophosphate, but a nucleoside diphosphate or nucleoside triphosphate could be considered a nucleotide as well.

Without the phosphate group, the nucleobase and sugar compose a nucleoside. The phosphate groups form bonds with either the 2, 3, or 5-carbon of the sugar, with the 5-carbon site most common. Cyclic nucleotides form when the phosphate group is bound to two of the sugar's hydroxyl groups. Nucleotides contain either a purine or a pyrimidine base. Ribonucleotides are nucleotides in which the sugar is ribose. Deoxyribonucleotidesare nucleotides in which the sugar is deoxyribose.

Nucleic acids are polymeric macromolecules made from nucleotide monomers. In DNA, the purine bases are adenine and guanine, while the pyrimidines are thymine and cytosine. RNA

uses uracil in place of thymine. Adenine always pairs with thymine by 2 hydrogen bonds, while guanine pairs with cytosine through 3 hydrogen bonds, in each case because of the unique structures of the bases.

Nucleosides

Nucleosides are glycosylamines that can be thought of as nucleotides without a phosphate group. A nucleoside consists simply of a nucleobase (also termed a nitrogenous base) and a 5-carbon sugar (either ribose or deoxyribose), whereas a nucleotide is composed of a nucleobase, a five-carbon sugar, and one or more phosphate groups. In a nucleoside, the base is bound to either ribose or deoxyribose via a beta-glycosidic linkage.

Examples of nucleosides include cytidine,uridine, adenosine, guanosine, thymidine and inosine.

While a nucleoside is a nucleobase linked to a sugar, a nucleotide is composed of a nucleoside *and* 1 or more phosphate groups. Thus, nucleosides can be phosphorylated by specific kinases in the cell on the sugar's primary alcohol group ($-CH_2-OH$) to produce nucleotides. Nucleotides are the molecular building-blocks of DNA and RNA.

Nucleosides can be produced by de novo synthesis pathways, in particular in the liver, but they are more abundantly supplied via ingestion and digestion of nucleic acids in the diet, whereby nucleotidases break down *nucleotides* (such as the thymidine monophosphate) into *nucleosides* (such as thymidine) and phosphate. The nucleosides, in turn, are subsequently broken down:

- in the lumen of the digestive system by nucleosidases into nucleobases and ribose or deoxyribose.

In addition, nucleotides can be broken down:

DNA structure

Deoxyribonucleic acid (**DNA**) is a molecule that carries the genetic instructions used in the growth, development, functioning and reproduction of all known living organisms and many viruses. DNA and RNA are nucleic acids; alongside proteins, lipids and complex carbohydrates (polysaccharides), they are one of the four major types ofmacromolecules that are essential for all known forms of life. Most DNA molecules consist of two biopolymer strands coiled around each other to form a double helix.

The two DNA strands are termed polynucleotides since they are composed of simpler monomer units called nucleotides.Each nucleotide is composed of one of four nitrogen-containing nucleobases—either cytosine (C), guanine (G), adenine (A), or thymine(T)—and a sugar called deoxyribose and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. The nitrogenous bases of the two separate polynucleotide strands are bound together (according to base pairing rules (A with T, and C with G) with hydrogen bonds to make double-stranded DNA. The total amount of related DNA base pairs on Earth is estimated at 5.0×10^{37}, and weighs 50 billion tonnes.In comparison, the total mass of the biosphere has been estimated to be as much as 4 trillion tons of carbon (TtC).DNA stores biological information. The DNA backbone is resistant to cleavage, and both strands of the double-stranded structure store the same biological information. This information is replicated as and when the two strands separate. A large part of DNA (more than 98% for humans) is non-coding, meaning that these sections do not serve as patterns for protein sequences.

The two strands of DNA run in opposite directions to each other and are thus antiparallel. Attached to each sugar is one of four types of nucleobases (informally, *bases*). It is the sequence of these four nucleobases along the backbone that encodes biological information. RNA strands are created using

DNA strands as a template in a process called transcription. Under the genetic code, these RNA strands are translated to specify the sequence of amino acids within proteins in a process called translation.

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

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

Non coding and Coding DNA

Noncoding DNA sequences are components of an organism's DNA that do not encode protein sequences. Some noncoding DNA is transcribed into functional non-coding RNA molecules (e.g. transfer RNA, ribosomal RNA, and regulatory RNAs). Other functions of noncoding DNA include the transcriptional and translational regulation of protein-coding sequences, scaffold attachment regions, origins of DNA replication, centromeres and telomeres.

The amount of noncoding DNA varies greatly among species. Where only a small percentage of the genome is responsible for coding proteins, the percentage of the genome performing regulatory functions is growing. When there is much non-coding DNA, a large proportion appears to have no biological function for the organism, as theoretically predicted in the 1960s. Since that time, this non-functional portion has often been referred to as "junk DNA", a term that has elicited strong responses over the years.[2]

When referring to DNA transcription, the **coding strand** is the DNA strand whose base sequence is corresponding to the base sequence of the RNA transcript produced (although with thymine replaced by uracil). It is this strand which contains codons, while the non-coding strand contains anti-codons. During transcription, RNA Pol II binds the non-coding strand, reads the anti-codons, and transcribes their sequence to synthesize an RNA transcript with complementary bases.

By convention, the coding strand is the strand used when displaying a DNA sequence. It is presented in the 5' to 3' direction.

The amount of total genomic DNA varies widely between organisms, and the proportion of coding and noncoding DNA within these genomes varies greatly as well. More than 98% of the human genome does not encode protein sequences, including most sequences within introns and most intergenic DNA. While overall genome size, and by extension the amount of noncoding DNA, are correlated to organism complexity, there are many exceptions. For example, the genome of the unicellular Polychaosdubium (formerly known as Amoeba dubia) has been reported to contain more than 200 times the amount of DNA in humans. The pufferfishTakifugurubripes genome is only about one eighth the size of the human genome, yet seems to have a comparable number of genes; approximately 90% of the Takifugu genome is noncoding DNA.

Repeated sequences

Repeated sequences are patterns of nucleic acids (DNA or RNA) that occur in multiple copies throughout the genome. Repetitive DNA was first detected because of its rapid reassociation kinetics.

In many organisms, a significant fraction of the genomic DNA is highly repetitive, with over two-thirds of the sequence consisting of repetitive elements in human. Debates regarding the potential in vivo functions of these elements have been long standing. Controversial references to 'junk' or 'selfish' DNA were put forward early on, implying that repetitive DNA segments are remainders from past evolution or autonomous self-replicating sequences hacking the cell machinery to proliferate. Repetitive elements found in eukaryotic genomes fall into different classes, depending on their mode of multiplication and/or structure. The disposition of repetitive elements consists either in arrays of tandemly repeated sequences, or in repeats dispersed throughout the genome (see below). Originally discovered by Barbara McClintock, dispersed repeats have been increasingly recognized as a potential source of genetic variation and regulation. Together with these regulatory roles, a structural role of repeated DNA in shaping the 3D folding of genomes has also been proposed. This hypothesis is only supported by a limited set of experimental evidence. For instance in human, mouse and fly, several classes of repetitive elements present a high tendency for co-localization within the nuclear space, suggesting that DNA repeats positions can be used by the cell as a genome folding map

There are 3 major categories of **repeated sequence** or **repeats**:

- Terminal repeats
- Tandem repeats: copies which lie adjacent to each other, either directly or inverted
    - Satellite DNA - typically found in centromeres and heterochromatin
    - Minisatellite - repeat units from about 10 to 60 base pairs, found in many places in the genome, including the centromeres
    - Microsatellite - repeat units of less than 10 base pairs; this includes telomeres, which typically have 6 to 8 base pair repeat units
- Interspersed repeats (aka. interspersed nuclear elements)
    - Transposable elements
        - DNA transposons
        - retrotransposons
            - LTR-retrotranposons (HERVs)
            - non LTR-retrotranposons
                - SINEs (**S**hort **I**nterspersed **N**uclear **E**lements)
                - LINEs (**L**ong **I**nterspersed **N**uclear **E**lements)
                - SVAs

Satellite DNA

Satellite DNA consists of very large arrays of tandemly repeating, non-coding DNA. Satellite DNA is the main component of functional centromeres, and form the main structural constituent of heterochromatin.

The name "satellite DNA" refers to how repetitions of a short DNA sequence tend to produce a different frequency of the nucleotides adenine, cytosine, guanine and thymine, and thus have a different density from bulk DNA - such that they form a second or 'satellite' band when genomic DNA is separated on a density gradient

Satellite DNA together with minisatellite microsatellite DNA, constitute the tendem repeats.

Minisatellite: is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 10-60 base pairs) are typically repeated 5-50 times. Minisatellites are notable for their high mutation rate and

high diversity in the population. Minisatellites are prominent in centromeres and telomers of chromosomes.

Microsatellites: Is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 2-5 base pairs) are typically repeated 5-50 times. Minisatellites are notable for their high mutation rate and high diversity in the population. Microsatellites often referred to as STRs (short tandem repeats)by forensic geneticists or as simple sequence repeats(SSRs) by plant geneticists. They are used in genetic linkage analysis/marker assisted selection to locate a gene or a mutation responsible for a given trait or disease.

Structure:

Satellite DNA adopts higher-order three-dimensional structures in eukaryotic organisms. This was demonstrated in the land crab *Gecarcinuslateralis*, whose DNA contains 3% of a GC-rich sequence consisting of repeats of a ~2100 base pair (bp) sequence called RU. The RU was arranged in long tandem arrays with approximately 16,000 copies per genome. Several RU sequences were cloned and sequenced to reveal conserved regions of conventional DNA sequences interspersed with four domains of microsatellite repeats biased in composition with purines on one strand and pyrimidines on the other, including mononucleotide repeats of G and C base pairs 20-25 bp in length. The most prevalent repeated sequences in the embedded microsatellite regions were CCT:AGG and CCCT:AGGG. The strand biased pyrimidine:purine repeating sequences were shown to adopt triple-stranded structures under superhelical stress or at slightly acidic pH.

Between the strand-biased microsatellite repeats and G:C mononucleotide repeats, all sequence variations retained one or two base pairs with A (purine) interrupting the pyrimidine-rich strand and T (pyrimidine) interrupting the purine-rich strand. This sequence feature adopted a highly distorted conformation as shown by its response to nuclease enzymes. The sequence TTAA was found in one variant of RU, and the strand-biased domain was subcloned and studied in greater detail.

A fifth region of the RU sequence was characterized by variations of a symmetrical DNA sequence of alternating purines and pyrimidines shown to adopt a left-handed Z-DNA helical structure in equilibrium with a stem-loop structure under superhelical stress. The sequence consisted of CGCAC:GTGCG and variations that retained the alternating purine and pyrimidine motif. A fragment containing the domain was excised and subcloned in order to examine structural properties of the alternating purine and pyrimidine motif independently of the four compositionally-biased repetitive sequences within RU. The palindromic sequence CGCACGTGCG:CGCACGTGCG, flanked by extended palindromic Z-DNA sequences over a 35 bp domain, adopted a Z-DNA structure with a symmetrical arrangement in equilibrium with a stem-loop structure centered on the palindrome containing the CGCAC:GTGCG motif. The CGCAC:GTGCG sequence was also found in tandem repeats with at least five copies immediately adjacent to one of the pyrimidine:purine biased divergent domains.

Conserved sequences showed virtually no differences among cloned RU sequences. Variations among cloned RU sequences were characterized by the number of microsatellite repeats, and also by the lengths of C and G stretches where triple stranded structures formed. Other regions of variability among cloned RU sequences were found adjacent to alternating purine and pyrimidine sequences with Z-DNA/stem-loop structures.

Tandem repeats

Tandem repeats occur in DNA when a pattern of one or more nucleotides is repeated and the repetitions are directly adjacent to each other. Several protein domains also form tandem repeats within their amino acid primary structure, such as Armadillo repeats. However, in proteins, perfect tandem

repeats are unlikely in most *in vivo* proteins, and most known repeats are in proteins which have been designed.

An example would be:

ATTCG ATTCGATTCG

in which the sequence ATTCG is repeated three times.

When between 10 and 60 nucleotides are repeated, it is called a minisatellite. Those with fewer are known as microsatellites or short tandem repeats.

When exactly two nucleotides are repeated, it is called a *dinucleotide repeat* (for example: ACACACAC…). The microsatellite instability in hereditary nonpolyposis colon cancer most commonly affects such regions.

When three nucleotides are repeated, it is called a *trinucleotide repeat* (for example: CAGCAGCAGCAG…), and abnormalities in such regions can give rise to trinucleotide repeat disorders.

When the repeat unit copy number is variable in the population being considered, it is called a variable number tandem repeat (VNTR). MeSH classifies variable number tandem repeats under minisatellites.

Tandem repeat describes a pattern that helps determine an individual's inherited traits

Uses:

Tandem repeats can be very useful in determining parentage. Short tandem repeats are used for certain genealogical DNA tests.

DNA is examined from *microsatellites* within the chromosomal DNA. Minisatellite is another way of saying special regions of the loci. Polymerase chain reaction (or PCR) is performed on the minisatellite areas. The PCR must be performed on each organism being tested. The amplified material is then run through electrophoresis. By checking the percentage of bands that match, parentage is determined.

Polymorphic tandem repeats (alias VNTRs) are also present in microorganisms and can be used to trace the origin of an outbreak. The corresponding assay in which a collection of VNTRs is typed to characterize a strain is most often called MLVA (Multiple Loci VNTR Analysis).

In the field of Computer Science, tandem repeats in strings (e.g., DNA sequences) can be efficiently detected using suffix trees or suffix arrays.

Studies in 2004 linked the unusual genetic plasticity of dogs to mutations in tandem repeats.

VNTRs

A **variable number tandem repeat** (or **VNTR**) is a location in a genome where a short nucleotide sequence is organized as a tandem repeat. These can be found on many chromosomes, and often show variations in length between individuals. Each variant acts as an inherited allele, allowing them to be used for personal or parental identification. Their analysis is useful ingenetics and biology research, forensics, and DNA fingerprinting.

There are two principal families of VNTRs: microsatellites and minisatellites. The former are repeats of sequences less than about 5 base pairs in length (an arbitrary cutoff), while the latter involve longer blocks. VNTRs with very short repeat blocks may be unstable - dinucleotide repeats may vary from one tissue to another within an individual, while trinucleotide repeats have been found to vary from one generation to another.

Uses

VNTRs were an important source of RFLP genetic markers used in linkage analysis (mapping) of diploid genomes. Now that many genomes have been sequenced, VNTRs have become essential to forensic crime investigations, via DNA fingerprintingand the CODIS database. When removed from surrounding DNA by the PCR or RFLP methods, and their size determined by gel electrophoresis or Southern blotting, they produce a pattern of bands unique to each individual. When tested with a group of independent VNTR markers, the likelihood of two unrelated individuals' having the same allelic pattern is extremely low. VNTR analysis is also being used to study genetic diversity and breeding patterns in populations of wild or domesticated animals. As such, VNTRs can be used to distinguish strains of bacterial pathogens. In this microbial forensics context, such assays are usually called Multiple Loci VNTR Analysis or MLVA.

Junk DNA

The term "junk DNA" became popular in the 1960s. According to T. Ryan Gregory, a genomic biologist, the first explicit discussion of the nature of junk DNA was done by David Comings in 1972 and he applied the term to all noncoding DNA. The term was formalized in 1972 by Susumu Ohno, who noted that the mutational load from deleterious mutations placed an upper limit on the number of functional loci that could be expected given a typical mutation rate. Ohno hypothesized that mammal genomes could not have more than 30,000 loci under selection before the "cost" from the mutational load would cause an inescapable decline in fitness, and eventually extinction. This prediction remains robust, with the human genome containing approximately 20,000 genes. Another source for Ohno's theory was the observation that even closely related species can have widely (orders-of-magnitude) different genome sizes, which had been dubbed the C-value paradox in 1971.

Though the fruitfulness of the term "junk DNA" has been questioned on the grounds that it provokes a strong a priori assumption of total non-functionality and though some have recommended using more neutral terminology such as "noncoding DNA" instead; "junk DNA" remains a label for the portions of a genome sequence for which no discernible function has been identified and that through comparative genomics analysis appear under no functional constraint suggesting that the sequence itself has provided no adaptive advantage. Since the late 70s it has become apparent that the majority of non-coding DNA in large genomes finds its origin in the selfish amplification of transposable elements, of which W. Ford Doolittle and Carmen Sapienza in 1980 wrote in the journal *Nature*: "When a given DNA, or class of DNAs, of unproven phenotypic function can be shown to have evolved a strategy (such as transposition) which ensures its genomic survival, then no other explanation for its existence is necessary." The amount of junk DNA can be expected to depend on the rate of amplification of these elements and the rate at which non-functional DNA is lost. In the same issue of *Nature.*

ENCODE suggested that over 80% of the human genome is biochemically functional has been criticized by other scientists, who argue that neither accessibility of segments of the genome to transcription factors nor their transcription guarantees that those segments have biochemical function and that their transcription is selectively advantageous.

Palindromes

A **palindromic sequence** is a nucleic acid sequence on double-stranded DNA or RNA wherein reading 5' (five-prime) to 3' (three prime) forward on one strand matches the sequence reading 5' to 3' on the complementary strand with which it forms a double helix. This definition of palindrome thus depends on complementary strands being palindromic of each other.

The meaning of palindrome in the context of genetics is slightly different from the definition used for words and sentences. Since a double helix is formed by two paired strands of nucleotides that run in

opposite directions in the 5'-to-3' sense, and the nucleotides always pair in the same way (Adenine (A) with Thymine (T) for DNA, with Uracil (U) for RNA; Cytosine (C) with Guanine (G)), a (single-stranded) nucleotide sequence is said to be a **palindrome** if it is equal to its reverse complement. For example, the DNA sequence ACCTAGGT is palindromic because its nucleotide-by-nucleotide complement is TGGATCCA, and reversing the order of the nucleotides in the complement gives the original sequence.

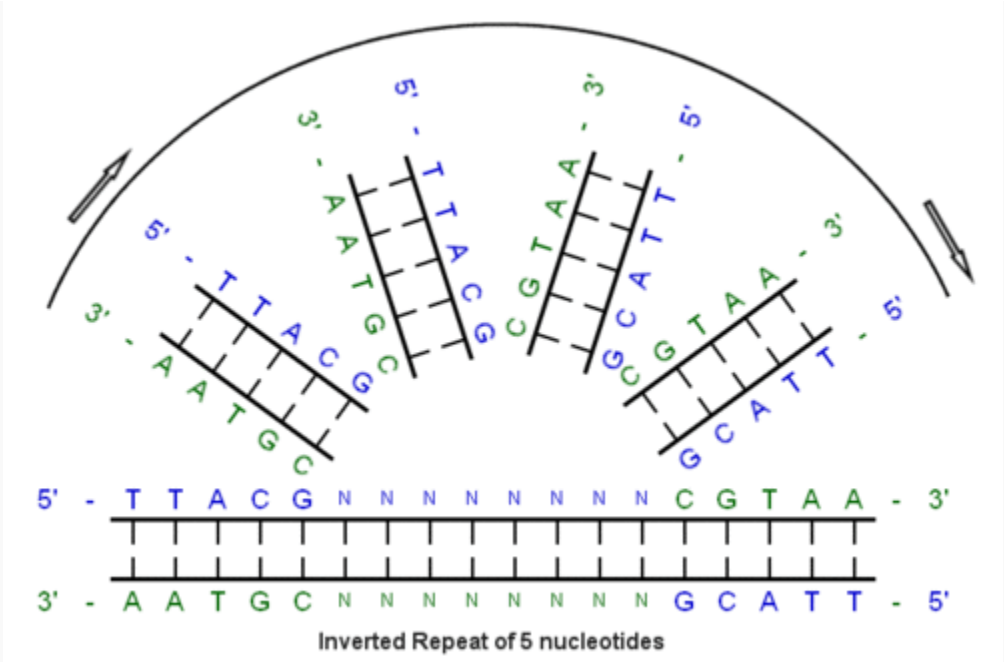
A palindromic nucleotide sequence can form a hairpin. Palindromic DNA motifs are found in most genomes or sets of genetic instructions. Palindromic motifs are made by the order of the nucleotides that specify the complex chemicals (proteins) which, as a result of those genetic instructions, the cell is to produce. They have been specially researched in bacterial chromosomes and in the so-called Bacterial Interspersed Mosaic Elements (BIMEs) scattered over them. Recently, a research genome sequencing project discovered that many of the bases on the Y chromosome are arranged as palindromes. A palindrome structure allows the Y chromosome to repair itself by bending over at the middle if one side is damaged.

Palindromes also appear to be found frequently in proteins, but their role in the protein function is not clearly known. It has recently been suggested that the existence of palindromes in peptides might be related to the prevalence of low-complexity regions in proteins, as palindromes are frequently associated with low-complexity sequences. Their prevalence might be also related to an alpha helical formation propensity of these sequences, or in formation of protein/protein complexes.

Inverted repeats:

An inverted repeat is a sequence of nucleotides followed downstream by its reverse complement.

Example of an inverted repeat



The 5 base-pair sequence on the left is "repeated" and "inverted" to form sequence on the right.

Inverted repeats are often described as "hotspots" of eukaryotic and prokaryotic genomic instability. Long inverted repeats are deemed to greatly influence the stability of the genome of various organisms. This is exemplified in *E. coli*, where genomic sequences with long inverted repeats are seldom replicated, but rather deleted with rapidity. Again, the long inverted repeats observed in yeast greatly favor recombination within the same and adjacent chromosomes, resulting in an equally very high rate of deletion. Finally, a very high rate of deletion and recombination were also observed in

mammalian chromosomes regions with inverted repeats. Reported differences in the stability of genomes of interrelated organisms are always an indication of a disparity in inverted repeats. The instability results from the tendency of inverted repeats to fold into hairpin- or cruciform-like DNA structures. These special structures can hinder or confuse DNA replication and other genomic activities. Thus, inverted repeats lead to special configurations in both RNA and DNA that can ultimately cause mutations and disease.

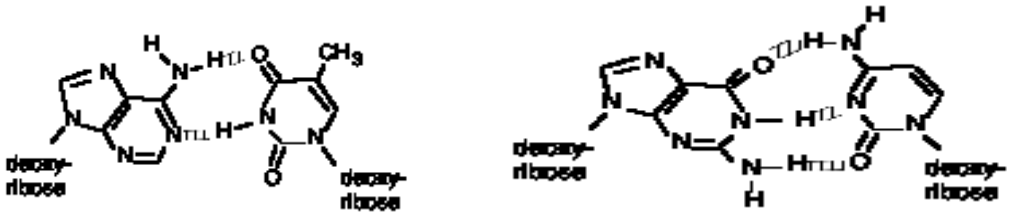
Conformations of DNA:

Three major forms of DNA are double stranded and connected by interactions between complementary base pairs. These are terms A-form, B-form, and Z-form DNA.

B-form DNA

The information from the base composition of DNA, the knowledge of dinucleotide structure, and the insight that the X-ray crystallography suggested a helical periodicity were combined by Watson and Crick in 1953 in their proposed model for a double helical structure for DNA. They proposed two strands of DNA -- each in a right-hand helix -- wound around the same axis. The two strands are held together by H-bonding between the bases (in anti conformation) as shown in Fig 1.

Major groove Major groove

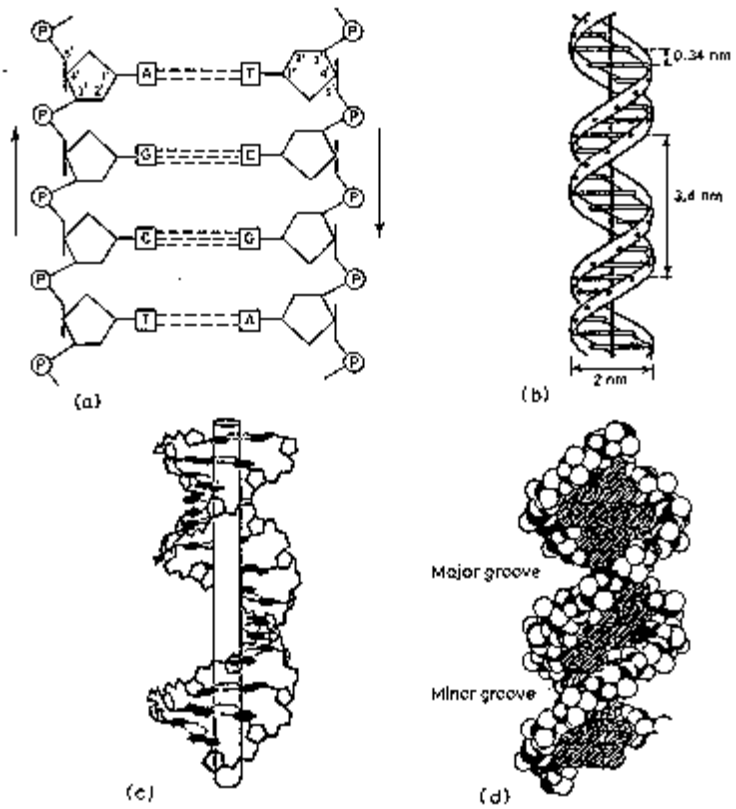


Minor groove Minor groove

Figure 1. (left) An A:T base pair and (right) a G:C base pair

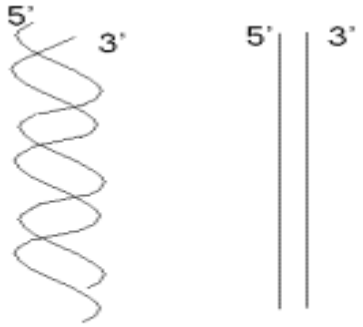
Bases fit in the double helical model if pyrimidine on one strand is always paired with purine on the other. From **Chargaff's rules**, the two strands will pair A with T and G with C. This pairs a keto base with an amino base, a purine with a pyrimidine. Two H-bonds can form between A and T, and three can form between G and C. This third H-bond in the G:C base pair is between the additional exocyclic amino group on G and the C2 keto group on C. The pyrimidine C2 keto group is not involved in hydrogen bonding in the A:T base pair.

These are the complementary base pairs. The base-pairing scheme immediately suggests a way to replicate and copy the the genetic information.



Antiparallel (a), plectonemically coiled (b, c, d) DNA strands. The arrows in a are pointed 3' to 5', but they illustrate the antiparallel nature of the duplex. The two strands of the duplex are antiparallel and plectonemically coiled. The nucleotides arrayed in a 5' to 3' orientation on one strand align with complementary nucleotides in the the 3' to 5' orientation of the opposite strand.

The two strands are not in a simple side-by-side arrangement, which would be called a paranemic joint (Fig. 2.). Rather the two strands are coiled around the same helical axis and are intertwined with themselves (which is referred to as a plectonemic coil). One consequence of this intertwining is that the two strands cannot be separated without the DNA rotating, one turn of the DNA for every "untwisting" of the two strands.



In a plectonemic coil, the two strands wrap around each other.
In a paranemic joint, the two strands align side-by-side.

Figure 2. Duplex DNA has the two strands wrapped around each other in a plectonemic coil (left), not a paranemic duplex (right).

A-form DNA

Three different forms of duplex nucleic acid have been described. The most common form, present in most DNA at neutral pH and physiological salt concentrations, is B-form. A thicker right-handed duplex with a shorter distance between the base pairs has been described for RNA-DNA duplexes and RNA-RNA duplexes. This is called A-form nucleic acid.

Z-form DNA

Z-DNA is a radically different duplex structure, with the two strands coiling in left-handed helices and a pronounced zig-zag (hence the name) pattern in the phosphodiester backbone. As previously mentioned, Z-DNA can form when the DNA is in an alternating purine-pyrimidine sequence such as GCGCGC, and indeed the G and C nucleotides are in different conformations, leading to the zig-zag pattern. The big difference is at the G nucleotide. It has the sugar in the C3' endoconformation (like A-form nucleic acid, and in contrast to B-form DNA) and the guanine base is in the synconformation. This places the guanine back over the sugar ring, in contrast to the usual anticonformation seen in A- and B-form nucleic acid. Note that having the base in the anticonformation places it in the position where it can readily form H-bonds with the complementary base on the opposite strand. The duplex in Z-DNA has to accomodate the distortion of this G nucleotide in the synconformation. The cytosine in the adjacent nucleotide of Z-DNA is in the "normal" C2' endo, anticonformation.

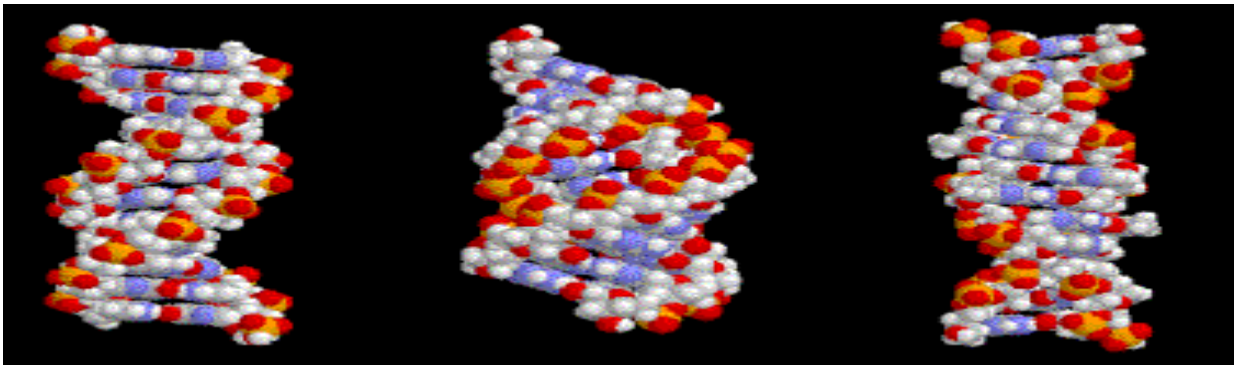


Figure. B-form (left), A-form (middle) and Z-DNA (right).

Supercoiling of DNA:

DNA supercoiling refers to the over- or under-winding of a DNA strand, and is an expression of the strain on that strand. Supercoiling is important in a number of biological processes, such as compacting DNA, and by regulating access to the genetic code, DNA supercoiling strongly affects DNA metabolism and possibly gene expression. Additionally, certain enzymes such as topoisomerases are able to change DNA topology to facilitate functions such as DNA replication or transcription. Mathematical expressions are used to describe supercoiling by comparing different coiled states to relaxed B-form DNA.

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

Positively supercoiled (overwound) DNA is transiently generated during DNA replication and transcription, and, if not promptly relaxed, inhibits (regulates) these processes. The simple figure eight is the simplest supercoil, and is the shape a circular DNA assumes to accommodate one too many or one too few helical twists. The two lobes of the figure eight will appear rotated either clockwise or counterclockwise with respect to one another, depending on whether the helix is over- or underwound. For each additional helical twist being accommodated, the lobes will show one more rotation about their axis. As a general rule, the DNA of most organisms is negatively supercoiled.

Lobal contortions of a circular DNA, such as the rotation of the figure-eight lobes above, are referred to as *writhe*. The above example illustrates that twist and writhe are interconvertible. Supercoiling can be represented mathematically by the sum of twist and writhe. The twist is the number of helical turns in the DNA and the writhe is the number of times the double helix crosses over on itself (these are the supercoils). Extra helical twists are positive and lead to positive supercoiling, while subtractive twisting causes negative supercoiling. Many topoisomerase enzymes sense supercoiling and either generate or dissipate it as they change DNA topology. DNA of most organisms is negatively supercoiled.

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

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

Functions:

DNA supercoiling is important for DNA packaging within all cells. Because the length of DNA can be thousands of times that of a cell, packaging this genetic material into the cell or nucleus (in eukaryotes) is a difficult feat. Supercoiling of DNA reduces the space and allows for DNA to be packaged.

DNA packaging is greatly increased during nuclear division events such as mitosis or meiosis, where DNA must be compacted and segregated to daughter cells

Supercoiling is also required for DNA/RNA synthesis. Because DNA must be unwound for DNA/RNA polymerase action, supercoils will result. The region ahead of the polymerase complex will be unwound; this stress is compensated with positive supercoils ahead of the complex.

RNA and its types:

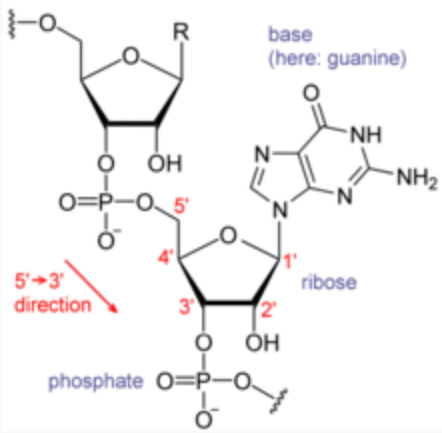
Ribonucleic acid (**RNA**) is a polymeric molecule essential in various biological roles in coding, decoding, regulation, and expression of genes. RNA and DNA are nucleic acids, and, along with proteins and carbohydrates, constitute the four major macromolecules essential for all known forms of life. Like DNA, RNA is assembled as a chain of nucleotides, but unlike DNA it is more often found in nature as a single-strand folded onto itself, rather than a paired double-strand. Cellular organisms use messenger RNA (***mRNA***) to convey genetic information (using the letters G, U, A, and C to denote the nitrogenous bases guanine, uracil, adenine, and cytosine) that directs synthesis of specific proteins. Many viruses encode their genetic information using an RNA genome.

Some RNA molecules play an active role within cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. One of these

active processes is protein synthesis, a universal function wherein mRNA molecules direct the assembly of proteins on ribosomes. This process uses transfer RNA (***tRNA***) molecules to deliver amino acids to the ribosome, where ribosomal RNA (***rRNA***) then links amino acids together to form proteins.

Structure:

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



Chemical structure of RNA

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

Types of RNA:

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

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

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

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

Functions:

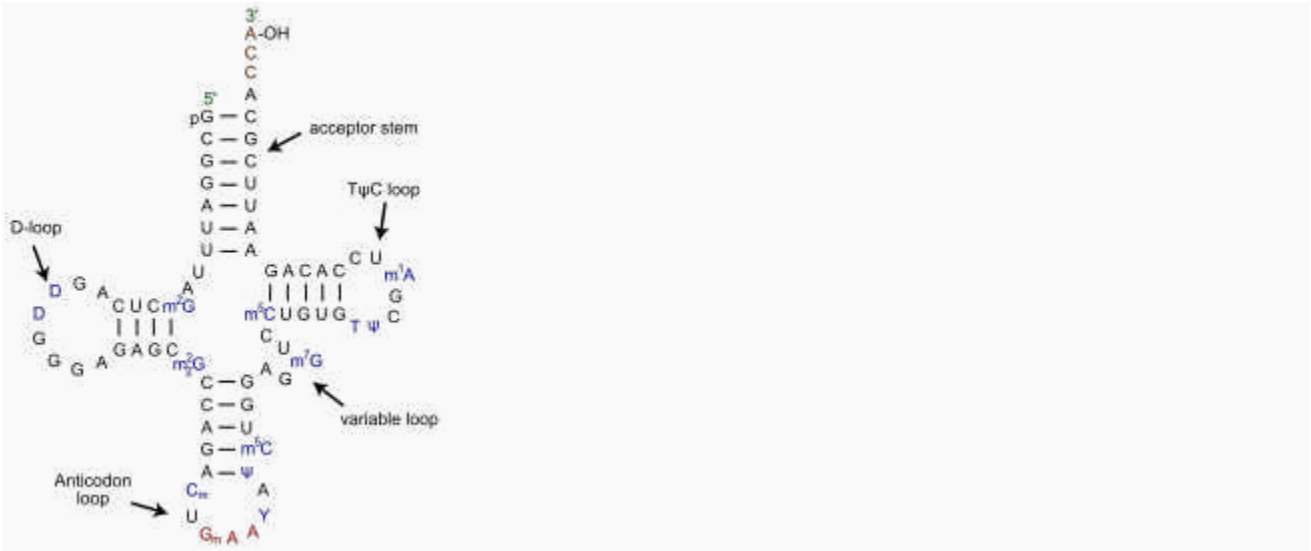
For many years, RNA was believed to have only three functions i-e mRNA, tRNA and rRNA.

In recent years RNA can avt as enzymes (ribozymes) which carries viral genetic information in cellular processor.

Structure of tRNA

A **transfer RNA** (abbreviated **tRNA** and formerly referred to as **sRNA**, for **soluble RNA**) is an adaptor molecule composed of RNA, typically 76 to 90 nucleotides in length, that serves as the physical link between the mRNA and the amino acid sequence of proteins. It does this by carrying an amino acid to the protein synthetic machinery of a cell (ribosome) as directed by a three-nucleotide sequence (codon) in amessenger RNA (mRNA). As such, tRNAs are a necessary component of translation, the biological synthesis of new proteins in accordance with the genetic code.

Structure:



The structure of tRNA can be decomposed into its primary structure, its secondary structure (usually visualized as the *cloverleaf structure*), and itstertiary structure (all tRNAs have a similar L-shaped 3D structure that allows them to fit into the P and A sites of the ribosome). The cloverleaf structure becomes the 3D L-shaped structure through coaxial stacking of the helices, which is a common RNA tertiary structure motif.

The lengths of each arm, as well as the loop 'diameter', in a tRNA molecule vary from species to species.

The tRNA structure consists of the following:

1. A 5'-terminal phosphate group.

2. The acceptor stem is a 7- to 9-base pair (bp) stem made by the base pairing of the 5'-terminal nucleotide with the 3'-terminal nucleotide (which contains the CCA 3'-terminal group used to attach the amino acid). The acceptor stem may contain non-Watson-Crick base pairs.
3. The CCA tail is a cytosine-cytosine-adenine sequence at the 3' end of the tRNA molecule. The amino acid loaded onto the tRNA by aminoacyltRNAsynthetases, to form aminoacyl-tRNA, is covalently bonded to the 3'-hydroxyl group on the CCA tail. This sequence is important for the recognition of tRNA by enzymes and critical in translation. In prokaryotes, the CCA sequence is transcribed in some tRNA sequences. In most prokaryotic tRNAs and eukaryotic tRNAs, the CCA sequence is added during processing and therefore does not appear in the tRNA gene.
4. The D arm is a 4- to 6-bp stem ending in a loop that often contains dihydrouridine.
5. The anticodon arm is a 5-bp stem whose loop contains the anticodon. The tRNA 5'-to-3' primary structure contains the anticodon but in reverse order, since 3'-to-5' directionality is required to read the mRNA from 5'-to-3'.
6. The Tarm is a 4- to 5- bp stem containing the sequence TΨC where Ψ is pseudouridine, a modified uridine.
7. Bases that have been modified, especially by methylation (e.g. tRNA (guanine-N7-)-methyltransferase), occur in several positions throughout the tRNA. The first anticodon base, or wobble-position, is sometimes modified to inosine (derived from adenine), pseudouridine or lysidine (derived from cytosine).

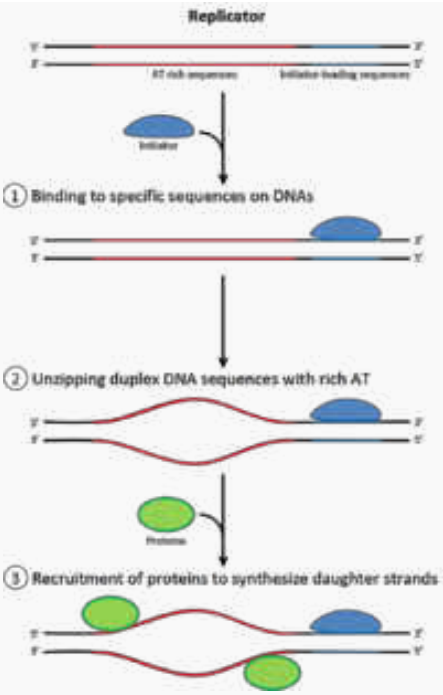
DNA replication

DNA replication is the biological process of producing two identical replicas of DNA from one original DNA molecule. This process occurs in all living organisms and is the basis for biological inheritance. DNA is made up of a double helix of two complementary strands. During replication, these strands are separated. Each strand of the original DNA molecule then serves as a template for the production of its counterpart, a process referred to as semi-conservative replication

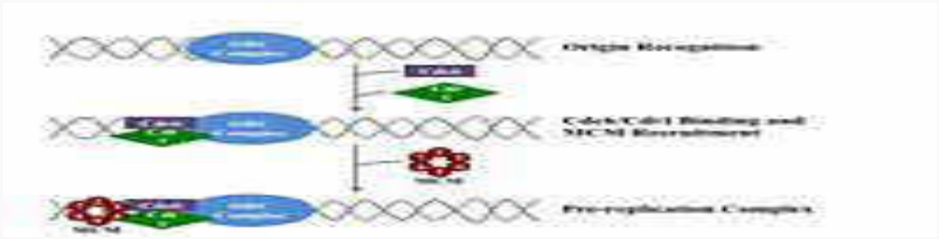
DNA replication begins at specific locations, or origins of replication, in the genome Unwinding of DNA at the origin and synthesis of new strands results in replication forks growing bi-directionally from the origin. A number of proteins are associated with the replication fork to help in the initiation and continuation of DNA synthesis. Most prominently, DNA polymerase synthesizes the new strands by adding nucleotides that complement each (template) strand. DNA replication occurs during the S-stage of interphase.

DNA replication, like all biological polymerization processes, proceeds in three enzymatically catalyzed and coordinated steps: initiation, elongation and termination.

Initiation



Role of initiators for initiation of DNA replication



Formation of pre-replication complex

For a cell to divide, it must first replicate its DNA This process is initiated at particular points in the DNA, known as "origins", which are targeted by initiator proteins. In *E. coli* this protein is DnaA; in yeast, this is the origin recognition complex Sequences used by initiator proteins tend to be "AT-rich" (rich in adenine and thymine bases), because A-T base pairs have two hydrogen bonds (rather than the three formed in a C-G pair) and thus are easier to strand separate Once the origin has been located, these initiators recruit other proteins and form the pre-replication complex, which unzips the double-stranded DNA.

Elongation

DNA polymerase has 5'-3' activity. All known DNA replication systems require a free 3' hydroxyl group before synthesis can be initiated. Four distinct mechanisms for DNA synthesis are recognized:

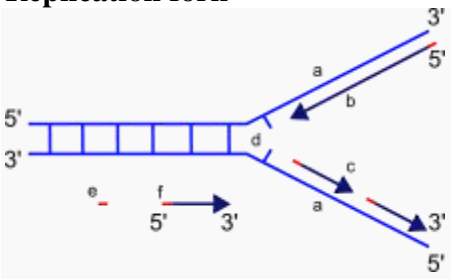
1. All cellular life forms and many DNA viruses, phages and plasmids use a primase to synthesize a short RNA primer with a free 3' OH group which is subsequently elongated by a DNA polymerase.
2. The retroelements (including retroviruses) employ a transfer RNA that primes DNA replication by providing a free 3′ OH that is used for elongation by the reverse transcriptase.
3. In the adenoviruses and the φ29 family of bacteriophages, the 3' OH group is provided by the side chain of an amino acid of the genome attached protein (the terminal protein) to which nucleotides are added by the DNA polymerase to form a new strand.

4. In the single stranded DNA viruses — a group that includes the circoviruses, the geminiviruses, the parvoviruses and others — and also the many phages and plasmids that use the rolling circle replication (RCR) mechanism, the RCR endonuclease creates a nick in the genome strand (single stranded viruses) or one of the DNA strands (plasmids). The 5′ end of the nicked strand is transferred to a tyrosine residue on the nuclease and the free 3′ OH group is then used by the DNA polymerase to synthesize the new strand.

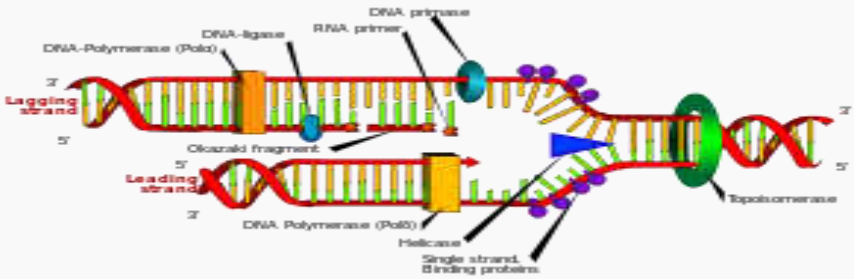
The first is the best known of these mechanisms and is used by the cellular organisms. In this mechanism, once the two strands are separated, primase adds RNA primers to the template strands. The leading strand receives one RNA primer while the lagging strand receives several. The leading strand is continuously extended from the primer by a DNA polymerase with high processivity, while the lagging strand is extended discontinuously from each primer forming Okazaki fragments. RNase removes the primer RNA fragments, and a low processivity DNA polymerase distinct from the replicative 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 a replication fork with two prongs. In bacteria, which have a single origin of replication on their circular chromosome, this process 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

Replication fork



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



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.

Termination

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 way. Because eukaryotes have linear chromosomes, DNA replication is unable to reach the very end of the chromosomes, but ends at the telomere region of repetitive DNA close to the ends. This shortens the telomere of the daughter DNA strand. Shortening of the telomeres is a normal process in somatic cells. As a result, cells can only divide 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. Increased telomerase activity is one of the hallmarks of cancer.

Termination requires that the progress of the DNA replication fork must stop or be blocked. 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* regulates 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.

Transcription

Transcription is the first step of gene expression, in which a particular segment of DNA is copied into RNA (mRNA) by the enzyme RNA polymerase. Both DNA and RNA are nucleic acids, which use base pairs of nucleotides as a complementary language. During transcription, a DNA sequence is read by an RNA polymerase, which produces a complementary, antiparallel RNA strand called a primary transcript.

Transcription proceeds in the following general steps:

1. RNA polymerase, together with one or more general transcription factors, binds to promoter DNA.
2. RNA polymerase creates a transcription bubble, which separates the two strands of the DNA helix. This is done by breaking the hydrogen bonds between complementary DNA nucleotides.
3. RNA polymerase adds RNA nucleotides (which are complementary to the nucleotides of one DNA strand).
4. RNA sugar-phosphate backbone forms with assistance from RNA polymerase to form an RNA strand.
5. Hydrogen bonds of the RNA–DNA helix break, freeing the newly synthesized RNA strand.
6. If the cell has a nucleus, the RNA may be further processed. This may include polyadenylation, capping, and splicing.

The RNA may remain in the nucleus or exit to the cytoplasm through the nuclear pore complex

Transcription is divided into *initiation*, *promoter escape*, *elongation* and *termination*.

Initiation

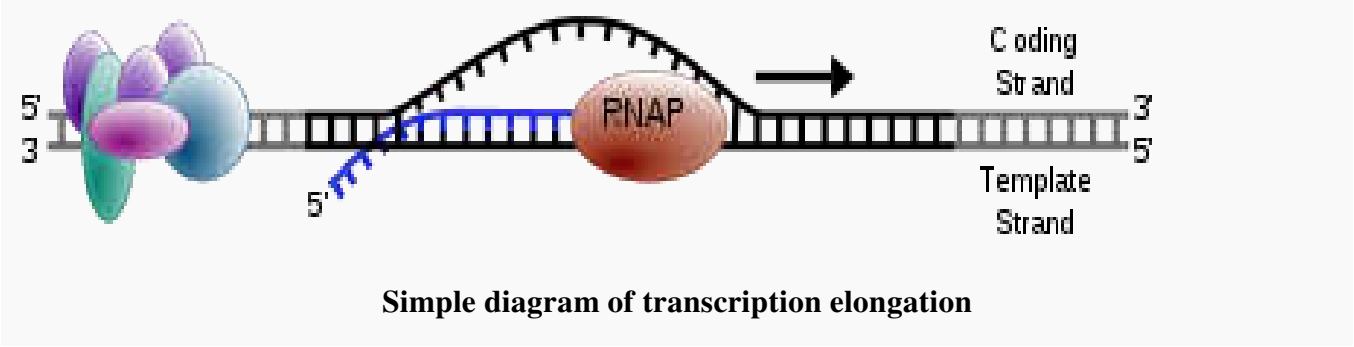
Transcription begins with the binding of RNA polymerase, together with one or more general transcription factor, to a specific DNA sequence referred to as a "promoter" to form an RNA polymerase-promoter "closed complex" (called a "closed complex" because the promoter DNA is fully double-stranded).

RNA polymerase, assisted by one or more general transcription factors, then unwinds approximately 14 base pairs of DNA to form an RNA polymerase-promoter "open complex" (called an

"open complex" because the promoter DNA is partly unwound and single-stranded) that contains an unwound, single-stranded DNA region of approximately 14 base pairs referred to as the "transcription bubble."

RNA polymerase, assisted by one or more general transcription factors, then selects a transcription start site in the transcription bubble, binds to an initiating NTP and an extending NTP (or a short RNA primer and an extending NTP) complementary to the transcription start site sequence, and catalyzes bond formation to yield an initial RNA product.

Elongation



Simple diagram of transcription elongation

One strand of the DNA, the *template strand* (or noncoding strand), is used as a template for RNA synthesis. As transcription proceeds, RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy. Although RNA polymerase traverses the template strand from 3' → 5', the coding (non-template) strand and newly formed RNA can also be used as reference points, so transcription can be described as occurring 5' → 3'. This produces an RNA molecule from 5' → 3', an exact copy of the coding strand (except that thyminesare replaced with uracils, and the nucleotides are composed of a ribose (5-carbon) sugar where DNA has deoxyribose (one fewer oxygen atom) in its sugar-phosphate backbone).

mRNA transcription can involve multiple RNA polymerases on a single DNA template and multiple rounds of transcription (amplification of particular mRNA), so many mRNA molecules can be rapidly produced from a single copy of a gene.

Termination

Transcription termination in **eukaryotes** is less understood but involves cleavage of the new transcript followed by template-independent addition of adenines at its new 3' end, in a process called polyadenylation.

Post transcriptional modifications

It is the process in eukaryotic cells where primary transcript RNA is converted into mature RNA. A notable example is the conversion of precursor messenger RNA into mature messenger RNA (mRNA) that occurs prior to protein translation. The process includes three major steps: addition of a 5' cap, addition of a 3' poly-adenylation tail, and splicing. This process is vital for the correct translation of the genomes of eukaryotes because the initial precursor mRNA produced during transcription contains both exons (coding or important sequences involved in translation), and introns (non-coding sequences).

The pre-mRNA molecule undergoes three main modifications. These modifications are5' capping, 3' polyadenylation, and RNA splicing, which occur in the cell nucleus before the RNA is translated.

5' Processing

5′ cap

Capping of the pre-mRNA involves the addition of 7-methylguanosine (mG) to the 5' end. To achieve this, the terminal 5' phosphate requires removal, which is done with the aid of a phosphatase enzyme. The enzyme guanosyltransferase then catalyses the reaction, which produces the diphosphate 5' end. The diphosphate 5' end then attacks the alpha phosphorus atom of a GTP molecule in order to add the guanine residue in a 5'5' triphosphate link. The enzyme (guanine-*N*-)-methyltransferase ("cap MTase") transfers a methyl group from S-adenosyl methionine to the guanine ring. This type of cap, with just the (mG) in position is called a cap 0 structure. The ribose of the adjacentnucleotide may also be methylated to give a cap 1. Methylation of nucleotides downstream of the RNA molecule produce cap 2, cap 3 structures and so on. In these cases the methyl groups are added to the 2' OH groups of the ribose sugar. The cap protects the 5' end of the primary RNA transcript from attack by ribonucleases that have specificity to the 3'5' phosphodiester bonds.

3' Processing

Cleavage and polyadenylation

The pre-mRNA processing at the 3' end of the RNA molecule involves cleavage of its 3' end and then the addition of about 250 adenine residues to form a poly(A) tail. The cleavage and adenylation reactions occur if a polyadenylation signal sequence (5'- AAUAAA-3') is located near the 3' end of the pre-mRNA molecule, which is followed by another sequence, which is usually **(5'-CA-3')** and is the site of cleavage. A **GU-rich sequence** is also usually present further downstream on the pre-mRNA molecule. After the synthesis of the sequence elements, two multisubunit proteins called cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CStF) are transferred from RNA Polymerase II to the RNA molecule. The two factors bind to the sequence elements. A protein complex forms that contains additional cleavage factors and the enzyme Polyadenylate Polymerase (PAP). This complex cleaves the RNA between the polyadenylation sequence and the GU-rich sequence at the cleavage site marked by the (5'-CA-3') sequences. Poly(A) polymerase then adds about 200 adenine units to the new 3' end of the RNA molecule using ATP as a precursor. As the poly(A) tail is synthesised, it binds multiple copies of poly(A) binding protein, which protects the 3'end from ribonuclease digestion.

Splicing

RNA splicing is the process by which introns, regions of RNA that do not code for protein, are removed from the pre-mRNA and the remaining exons connected to re-form a single continuous molecule. Although most RNA splicing occurs after the complete synthesis and end-capping of the pre-mRNA, transcripts with many exons can be spliced co-transcriptionally. The splicing reaction is catalyzed by a large protein complex called the spliceosome assembled from proteins and small nuclear RNA molecules that recognize splice sites in the pre-mRNA sequence. Many pre-mRNAs, including those encoding antibodies, can be spliced in multiple ways to produce different mature mRNAs that encode different protein sequences. This process is known as alternative splicing, and allows production of a large variety of proteins from a limited amount of DNA.

Translation

Translation is the process in which cellular ribosomes create proteins.

In translation, messenger RNA (mRNA)-produced by transcription from DNA-is decoded by a ribosome to produce a specific amino acid chain, or polypeptide. The polypeptide later folds into an active protein and performs its functions in the cell. The ribosome facilitates decoding by inducing the binding of complementary tRNA anticodon sequences to mRNAcodons. The tRNAs carry specific

amino acids that are chained together into a polypeptide as the mRNA passes through and is "read" by the ribosome. The entire process is a part of gene expression.

In brief, translation proceeds in three phases:

1. **Initiation**: The ribosome assembles around the target mRNA. The first tRNA is attached at the start codon.
2. **Elongation**: The tRNA transfers an amino acid to the tRNA corresponding to the next codon. The ribosome then moves (*translocates)* to the next mRNA codon to continue the process, creating an amino acid chain.
3. **Termination**: When a stop codon is reached, the ribosome releases the polypeptide.

In bacteria, translation occurs in the cell's cytoplasm, where the large and small subunits of the ribosome bind to the mRNA. In eukaryotes, translation occurs in the cytosol or across the membrane of the endoplasmic reticulum in a process calledvectorial synthesis. In many instances, the entire ribosome/mRNA complex binds to the outer membrane of the rough endoplasmic reticulum (ER); the newly created polypeptide is stored inside the ER for later vesicle transport and secretionoutside of the cell.

Many types of transcribed RNA, such as transfer RNA, ribosomal RNA, and small nuclear RNA, do not undergo translation into proteins.

A number of antibiotics act by inhibiting translation. These include anisomycin, cycloheximide, chloramphenicol, tetracycline, streptomycin, erythromycin, and puromycin. Prokaryotic ribosomes have a different structure from that of eukaryotic ribosomes, and thus antibiotics can specifically target bacterial infections without any harm to a eukaryotic host's cells.

Process:

The basic process of protein production is addition of one amino acid at a time to the end of a protein. This operation is performed by a ribosome. The choice of amino acid type to add is determined by an mRNA molecule. Each amino acid added is matched to a three nucleotide subsequence of the mRNA. For each such triplet possible, the corresponding amino acid is accepted. The successive amino acids added to the chain are matched to successive nucleotide triplets in the mRNA. In this way the sequence of nucleotides in the template mRNA chain determines the sequence of amino acids in the generated amino acid chain. Addition of an amino acid occurs at the C-terminus of the peptide and thus translation is said to be amino-to-carboxyl directed.

The mRNA carries genetic information encoded as a ribonucleotide sequence from the chromosomes to the ribosomes. The ribonucleotides are "read" by translational machinery in a sequence of nucleotide triplets called codons. Each of those triplets codes for a specific amino acid.

The ribosome molecules translate this code to a specific sequence of amino acids. The ribosome is a multisubunit structure containing rRNA and proteins. It is the "factory" where amino acids are assembled into proteins. tRNAs are small noncoding RNA chains (74-93 nucleotides) that transport amino acids to the ribosome. tRNAs have a site for amino acid attachment, and a site called an anticodon. The anticodon is an RNA triplet complementary to the mRNA triplet that codes for their cargo amino acid.

AminoacyltRNAsynthetases (enzymes) catalyze the bonding between specific tRNAs and the amino acids that their anticodon sequences call for. The product of this reaction is an aminoacyl-tRNA. This aminoacyl-tRNA is carried to the ribosome by EF-Tu, where mRNA codons are matched through complementary base pairing to specific tRNA anticodons. Aminoacyl-tRNAsynthetases that mispairtRNAs with the wrong amino acids can produce mischarged aminoacyl-tRNAs, which can result in inappropriate amino acids at the respective position in protein.

The ribosome has three sites for tRNA to bind. They are the aminoacyl site (abbreviated A), the peptidyl site (abbreviated P) and the exit site (abbreviated E). With respect to the mRNA, the three sites are oriented 5’ to 3’ E-P-A, because ribosomes move toward the 3' end of mRNA. The A site binds the incoming tRNA with the complementary codon on the mRNA. The P site holds the tRNA with the growing polypeptide chain. The E site holds the tRNA without its amino acid. When an aminoacyl-tRNA initially binds to its corresponding codon on the mRNA, it is in the A site. Then, a peptide bond forms between the amino acid of the tRNA in the A site and the amino acid of the charged tRNA in the P site. The growing polypeptide chain is transferred to the tRNA in the A site. Translocation occurs, moving the tRNA in the P site, now without an amino acid, to the E site; the tRNA that was in the A site, now charged with the polypeptide chain, is moved to the P site. The tRNA in the E site leaves and another aminoacyl-tRNA enters the A site to repeat the process.

After the new amino acid is added to the chain, and after the mRNA is released out of the nucleus and into the ribosome's core. The energy required for translation of proteins is significant.

In activation, the correct amino acid is covalently bonded to the correct transfer RNA (tRNA). Termination of the polypeptide happens when the A site of the ribosome faces a stop codon (UAA, UAG, or UGA). No tRNA can recognize or bind to this codon. Instead, the stop codon induces the binding of a release factor protein that prompts the disassembly of the entire ribosome/mRNA complex.

The process of translation is highly regulated in both eukaryotic and prokaryotic organisms. Regulation of translation can impact the global rate of protein synthesis which is closely coupled to the metabolic and proliferative state of a cell. In addition, recent work has revealed that genetic differences and their subsequent expression as mRNAs can also impact translation rate in an RNA-specific manner.