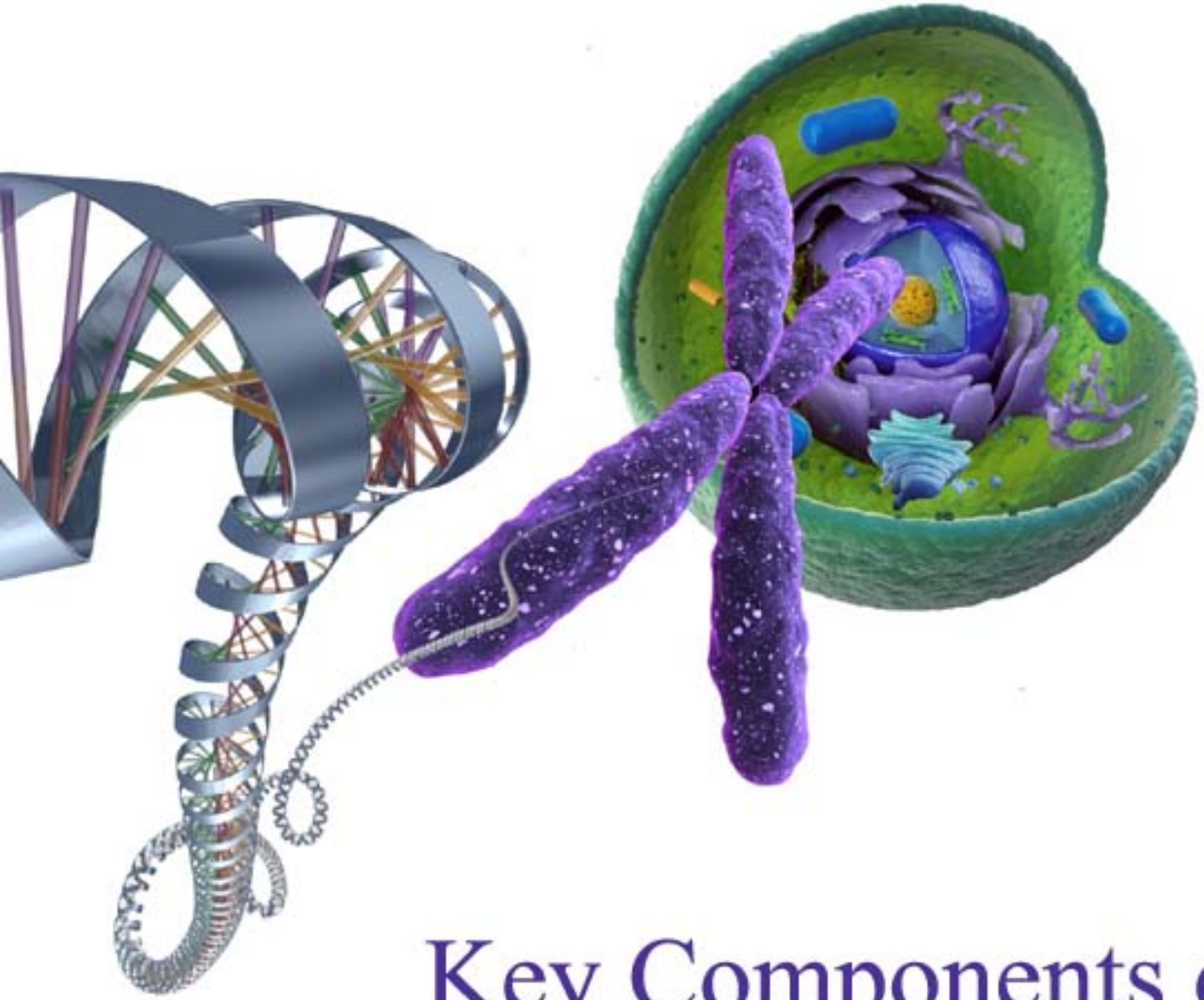


Key Components of Genetics

Lorena Shuman

First Edition, 2012

ISBN 978-81-323-3322-7

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

Table of Contents

Chapter- 1

Chromosome

A **chromosome** is an organized structure of DNA and protein that is found in cells. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences. Chromosomes also contain DNA-bound proteins, which serve to package the DNA and control its functions. The word *chromosome* comes from the Greek χρῶμα (*chroma*, colour) and σῶμα (*soma*, body) due to their property of being very strongly stained by particular dyes.

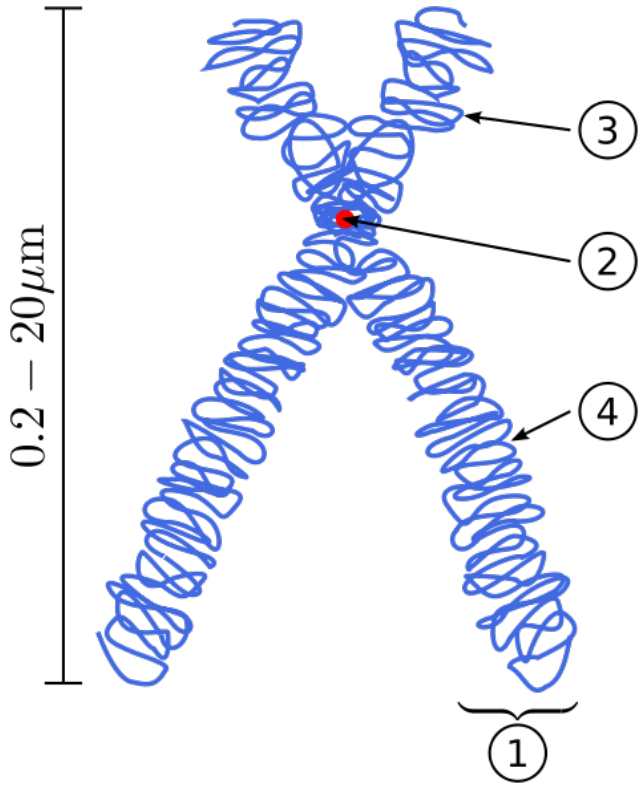


Diagram of a replicated and condensed metaphase eukaryotic chromosome. (1) Chromatid – one of the two identical parts of the chromosome after S phase. (2) Centromere – the point where the two chromatids touch, and where the microtubules attach. (3) Short arm. (4) Long arm.

Chromosomes vary widely between different organisms. The DNA molecule may be circular or linear, and can be composed of 10,000 to 1,000,000,000 nucleotides in a long chain. Typically eukaryotic cells (cells with nuclei) have large linear chromosomes and prokaryotic cells (cells without defined nuclei) have smaller circular chromosomes, although there are many exceptions to this rule. Furthermore, cells may contain more than one type of chromosome; for example, mitochondria in most eukaryotes and chloroplasts in plants have their own small chromosomes.

In eukaryotes, nuclear chromosomes are packaged by proteins into a condensed structure called chromatin. This allows the very long DNA molecules to fit into the cell nucleus. The structure of chromosomes and chromatin varies through the cell cycle. Chromosomes are the essential unit for cellular division and must be replicated, divided, and passed successfully to their daughter cells so as to ensure the genetic diversity and survival of their progeny. Chromosomes may exist as either duplicated or unduplicated—unduplicated chromosomes are single linear strands, whereas duplicated chromosomes (copied during synthesis phase) contain two copies joined by a centromere.

Compaction of the duplicated chromosomes during mitosis and meiosis results in the classic four-arm structure (pictured to the right). Chromosomal recombination plays a vital role in genetic diversity. If these structures are manipulated incorrectly, through processes known as chromosomal instability and translocation, the cell may undergo mitotic catastrophe and die, or it may unexpectedly evade apoptosis leading to the progression of cancer.

In practice "chromosome" is a rather loosely defined term. In prokaryotes and viruses, the term genophore is more appropriate when no chromatin is present. However, a large body of work uses the term chromosome regardless of chromatin content. In prokaryotes DNA is usually arranged as a circle, which is tightly coiled in on itself, sometimes accompanied by one or more smaller, circular DNA molecules called plasmids. These small circular genomes are also found in mitochondria and chloroplasts, reflecting their bacterial origins. The simplest genophores are found in viruses: these DNA or RNA molecules are short linear or circular genophores that often lack structural proteins.

History

Chromosomes as vectors of heredity

In a series of experiments, Theodor Boveri gave the definitive demonstration that chromosomes are the vectors of heredity. His two principles were based upon the *continuity* of chromosomes and the *individuality* of chromosomes. It is the second of these principles that was so original. Boveri was able to test the proposal put forward by Wilhelm Roux, that each chromosome carries a different genetic load, and showed that Roux was right. Upon the rediscovery of Mendel, Boveri was able to point out the connection between the rules of inheritance and the behaviour of the chromosomes. It is interesting to see that Boveri influenced two generations of American cytologists:

Edmund Beecher Wilson, Walter Sutton and Theophilus Painter were all influenced by Boveri (Wilson and Painter actually worked with him).

In his famous textbook *The Cell*, Wilson linked Boveri and Sutton together by the Boveri-Sutton theory. Mayr remarks that the theory was hotly contested by some famous geneticists: William Bateson, Wilhelm Johannsen, Richard Goldschmidt and T.H. Morgan, all of a rather dogmatic turn-of-mind. Eventually complete proof came from chromosome maps in Morgan's own lab.

Chromosomes in eukaryotes

Eukaryotes (cells with nuclei such as those found in plants, yeast, and animals) possess multiple large linear chromosomes contained in the cell's nucleus. Each chromosome has one centromere, with one or two arms projecting from the centromere, although, under most circumstances, these arms are not visible as such. In addition, most eukaryotes have a small circular mitochondrial genome, and some eukaryotes may have additional small circular or linear cytoplasmic chromosomes.

In the nuclear chromosomes of eukaryotes, the uncondensed DNA exists in a semi-ordered structure, where it is wrapped around histones (structural proteins), forming a composite material called chromatin.

Chromatin

Chromatin is the complex of DNA and protein found in the eukaryotic nucleus, which packages chromosomes. The structure of chromatin varies significantly between different stages of the cell cycle, according to the requirements of the DNA.

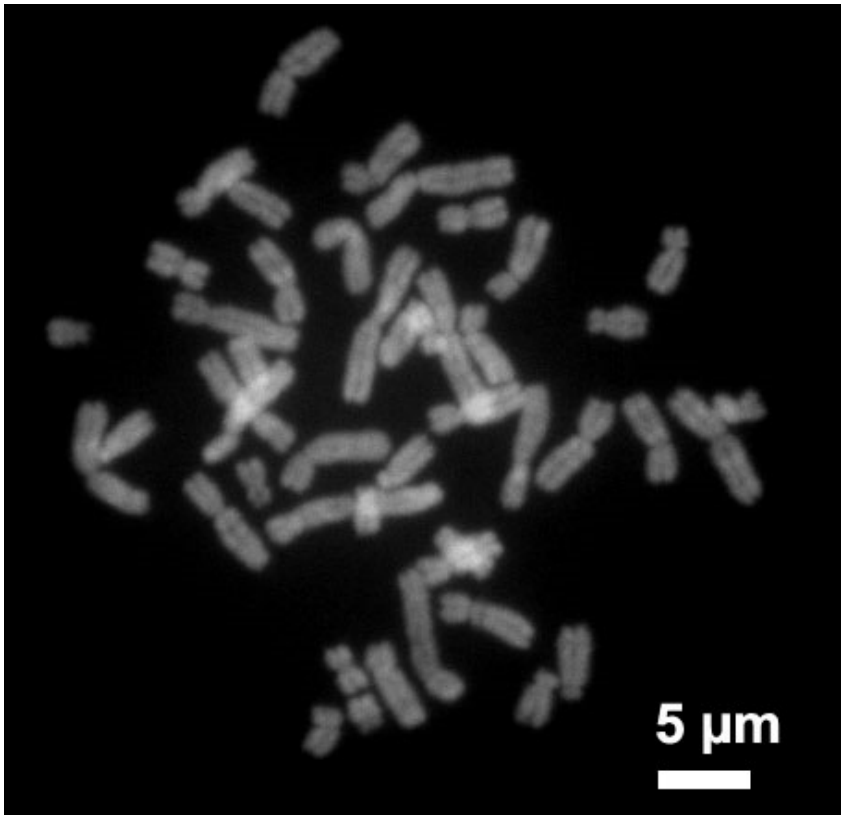
Interphase chromatin

During interphase (the period of the cell cycle where the cell is not dividing), two types of chromatin can be distinguished:

- Euchromatin, which consists of DNA that is active, e.g., being expressed as protein.
- Heterochromatin, which consists of mostly inactive DNA. It seems to serve structural purposes during the chromosomal stages. Heterochromatin can be further distinguished into two types:
    - *Constitutive heterochromatin*, which is never expressed. It is located around the centromere and usually contains repetitive sequences.
    - *Facultative heterochromatin*, which is sometimes expressed.

Individual chromosomes cannot be distinguished at this stage – they appear in the nucleus as a homogeneous tangled mix of DNA and protein.

Metaphase chromatin and division



Human chromosomes during metaphase

In the early stages of mitosis or meiosis (cell division), the chromatin strands become more and more condensed. They cease to function as accessible genetic material (transcription stops) and become a compact transportable form. This compact form makes the individual chromosomes visible, and they form the classic four arm structure, a pair of sister chromatids attached to each other at the centromere. The shorter arms are called *p arms* (from the French *petit*, small) and the longer arms are called *q arms* (*q* follows *p* in the Latin alphabet). This is the only natural context in which individual chromosomes are visible with an optical microscope.

During divisions, long microtubules attach to the centromere and the two opposite ends of the cell. The microtubules then pull the chromatids apart, so that each daughter cell inherits one set of chromatids. Once the cells have divided, the chromatids are uncoiled and can function again as chromatin. In spite of their appearance, chromosomes are structurally highly condensed, which enables these giant DNA structures to be contained within a cell nucleus (Fig. 2).

The self-assembled microtubules form the spindle, which attaches to chromosomes at specialized structures called kinetochores, one of which is present on each sister chromatid. A special DNA base sequence in the region of the kinetochores provides, along with special proteins, longer-lasting attachment in this region.

Chromosomes in prokaryotes

The prokaryotes – bacteria and archaea – typically have a single circular chromosome, but many variations do exist. Most bacteria have a single circular chromosome that can range in size from only 160,000 base pairs in the endosymbiotic bacterium *Candidatus Carsonella ruddii*, to 12,200,000 base pairs in the soil-dwelling bacterium *Sorangium cellulosum*. Spirochaetes of the genus *Borrelia* are a notable exception to this arrangement, with bacteria such as *Borrelia burgdorferi*, the cause of Lyme disease, containing a single linear chromosome.

Structure in sequences

Prokaryotic chromosomes have less sequence-based structure than eukaryotes. Bacteria typically have a single point (the origin of replication) from which replication starts, whereas some archaea contain multiple replication origins. The genes in prokaryotes are often organized in operons, and do not usually contain introns, unlike eukaryotes.

DNA packaging

Prokaryotes do not possess nuclei. Instead, their DNA is organized into a structure called the nucleoid. The nucleoid is a distinct structure and occupies a defined region of the bacterial cell. This structure is, however, dynamic and is maintained and remodeled by the actions of a range of histone-like proteins, which associate with the bacterial chromosome. In archaea, the DNA in chromosomes is even more organized, with the DNA packaged within structures similar to eukaryotic nucleosomes.

Bacterial chromosomes tend to be tethered to the plasma membrane of the bacteria. In molecular biology application, this allows for its isolation from plasmid DNA by centrifugation of lysed bacteria and pelleting of the membranes (and the attached DNA).

Prokaryotic chromosomes and plasmids are, like eukaryotic DNA, generally supercoiled. The DNA must first be released into its relaxed state for access for transcription, regulation, and replication.

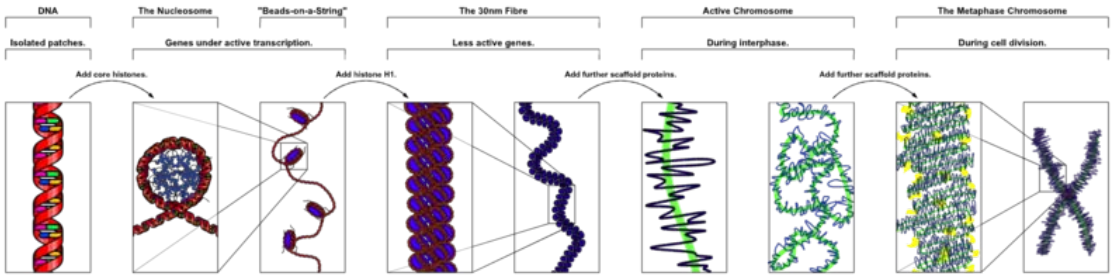


Fig. 2: The major structures in DNA compaction; DNA, the nucleosome, the 10nm "beads-on-a-string" fibre, the 30nm fibre and the metaphase chromosome.

Number of chromosomes in various organisms

Eukaryotes

These tables give the total number of chromosomes (including sex chromosomes) in a cell nucleus. For example, human cells are diploid and have 22 different types of autosome, each present as two copies, and two sex chromosomes. This gives 46 chromosomes in total. Other organisms have more than two copies of their chromosomes, such as bread wheat, which is *hexaploid* and has six copies of seven different chromosomes – 42 chromosomes in total.

Chromosome numbers in some plants

Plant Species	**#**
Arabidopsis thaliana (diploid)	10
Rye (diploid)	14
Maize (diploid or palaeotetraploid)	20
Einkorn wheat (diploid)	14
Durum wheat (tetraploid)	28
Bread wheat (hexaploid)	42
Cultivated tobacco (tetraploid)	48
Adder's Tongue Fern (diploid)	}

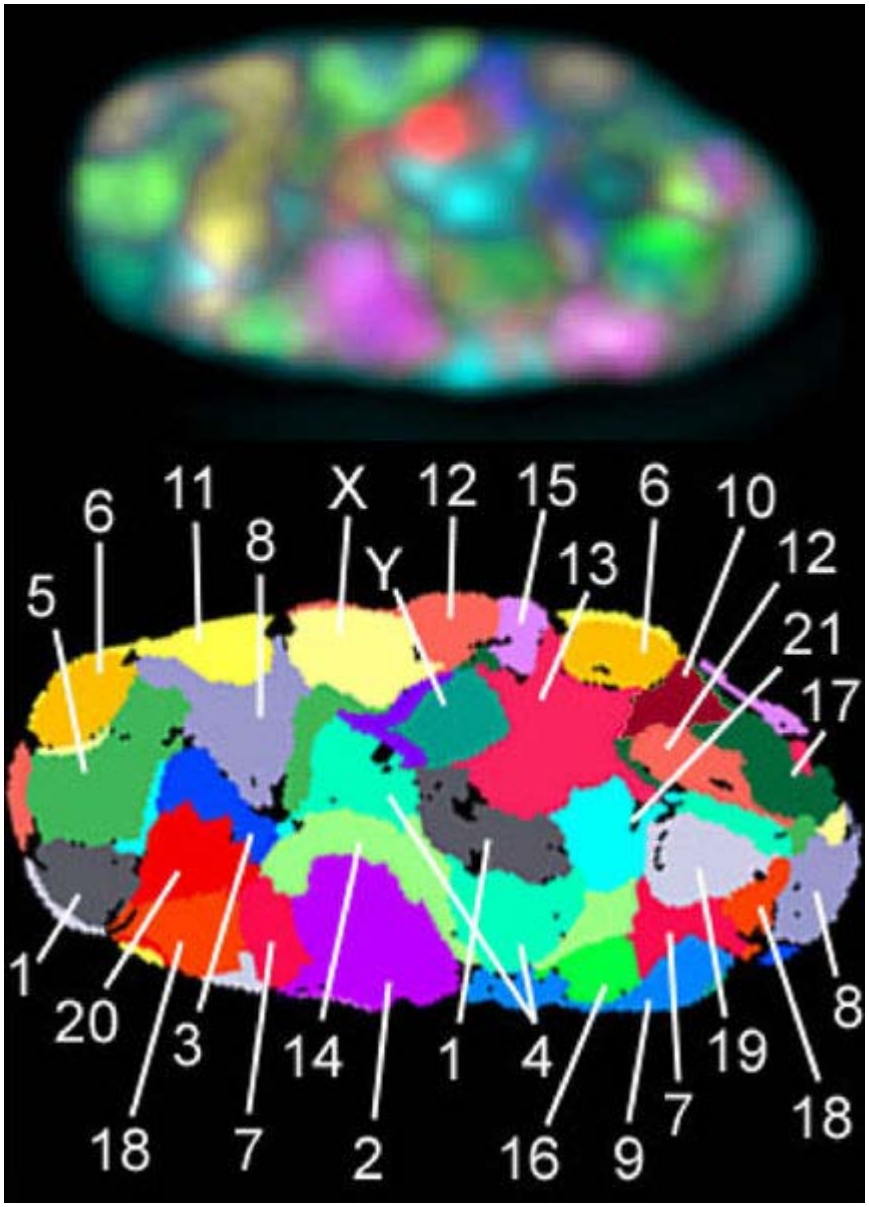
Chromosome numbers (2n) in some animals

Species	#	Species	#
Common fruit fly	8	Guinea Pig	64
Guppy (*poecilia reticulata*)	46	Garden snail	54
Earthworm (*Octodrilus complanatus*)	36	Tibetan fox	36
Domestic cat	38	Domestic pig	38
Laboratory mouse	40	Laboratory rat	42
Rabbit (*Oryctolagus cuniculus*)	44	Syrian hamster	44
Hares	48	Human	46
Gorillas, Chimpanzees	48	Domestic sheep	54
Elephants	56	Cow	60
Donkey	62	Horse	64
Dog	78	Kingfisher	132
Goldfish	100-104	Silkworm	56

Chromosome numbers in other organisms

Species	Large Chromosomes	Intermediate Chromosomes	Microchromosomes
Trypanosoma brucei	11	6	~100
Domestic Pigeon (*Columba livia domestics*)	18	-	59-63
Chicken	8	2 sex chromosomes	60

Normal members of a particular eukaryotic species all have the same number of nuclear chromosomes (see the table). Other eukaryotic chromosomes, i.e., mitochondrial and plasmid-like small chromosomes, are much more variable in number, and there may be thousands of copies per cell.



The 23 human chromosome territories during prometaphase in fibroblast cells

Asexually reproducing species have one set of chromosomes, which are the same in all body cells. However, asexual species can be either haploid or diploid.

Sexually reproducing species have somatic cells (body cells), which are diploid [2n] having two sets of chromosomes, one from the mother and one from the father. Gametes, reproductive cells, are haploid [n]: They have one set of chromosomes. Gametes are produced by meiosis of a diploid germ line cell. During meiosis, the matching chromosomes of father and mother can exchange small parts of themselves (crossover), and thus create new chromosomes that are not inherited solely from either parent. When a male and a female gamete merge (fertilization), a new diploid organism is formed.

Some animal and plant species are polyploid [Xn]: They have more than two sets of homologous chromosomes. Plants important in agriculture such as tobacco or wheat are often polyploid, compared to their ancestral species. Wheat has a haploid number of seven chromosomes, still seen in some cultivars as well as the wild progenitors. The more-common pasta and bread wheats are polyploid, having 28 (tetraploid) and 42 (hexaploid) chromosomes, compared to the 14 (diploid) chromosomes in the wild wheat.

Prokaryotes

Prokaryote species generally have one copy of each major chromosome, but most cells can easily survive with multiple copies. For example, *Buchnera*, a symbiont of aphids has multiple copies of its chromosome, ranging from 10–400 copies per cell. However, in some large bacteria, such as *Epulopiscium fishelsoni* up to 100,000 copies of the chromosome can be present. Plasmids and plasmid-like small chromosomes are, as in eukaryotes, very variable in copy number. The number of plasmids in the cell is almost entirely determined by the rate of division of the plasmid – fast division causes high copy number, and vice versa.

Karyotype

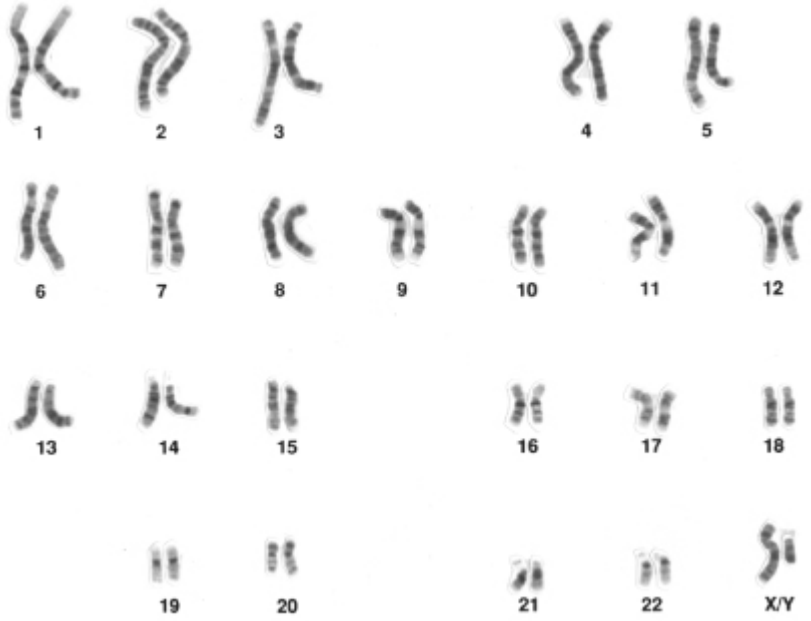


Figure 3: Karyogram of a human male

In general, the **karyotype** is the characteristic chromosome complement of a eukaryote species. The preparation and study of karyotypes is part of cytogenetics.

Although the replication and transcription of DNA is highly standardized in eukaryotes, *the same cannot be said for their karyotypes*, which are often highly variable. There may

be variation between species in chromosome number and in detailed organization. In some cases, there is significant variation within species. Often there is:

1. variation between the two sexes
2. variation between the germ-line and soma (between gametes and the rest of the body)
3. variation between members of a population, due to balanced genetic polymorphism
4. geographical variation between races
5. mosaics or otherwise abnormal individuals.

Also, variation in karyotype may occur during development from the fertilised egg.

The technique of determining the karyotype is usually called *karyotyping*. Cells can be locked part-way through division (in metaphase) in vitro (in a reaction vial) with colchicine. These cells are then stained, photographed, and arranged into a *karyogram*, with the set of chromosomes arranged, autosomes in order of length, and sex chromosomes (here X/Y) at the end: Fig. 3.

Like many sexually reproducing species, humans have special gonosomes (sex chromosomes, in contrast to autosomes). These are XX in females and XY in males.

Historical note

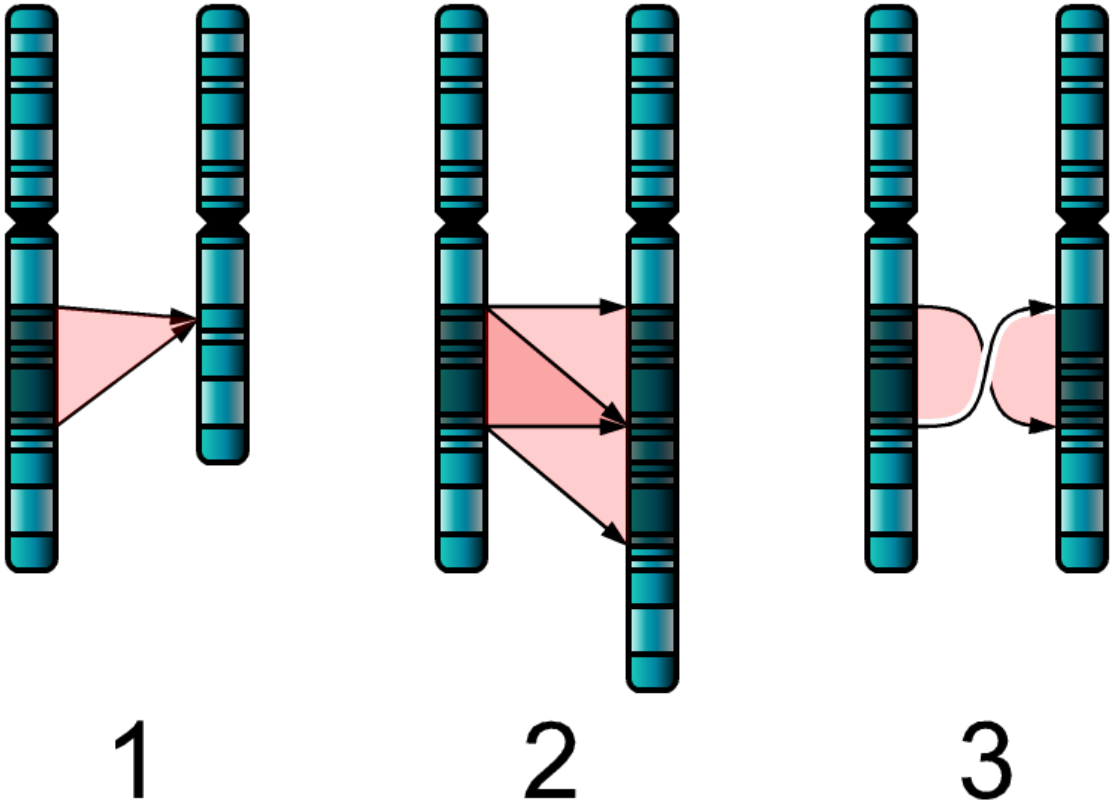
Investigation into the human karyotype took many years to settle the most basic question. How many chromosomes does a normal diploid human cell contain? In 1912, Hans von Winiwarter reported 47 chromosomes in spermatogonia and 48 in oogonia, concluding an XX/XO sex determination mechanism. Painter in 1922 was not certain whether the diploid number of man is 46 or 48, at first favouring 46. He revised his opinion later from 46 to 48, and he correctly insisted on humans having an XX/XY system.

New techniques were needed to definitively solve the problem:

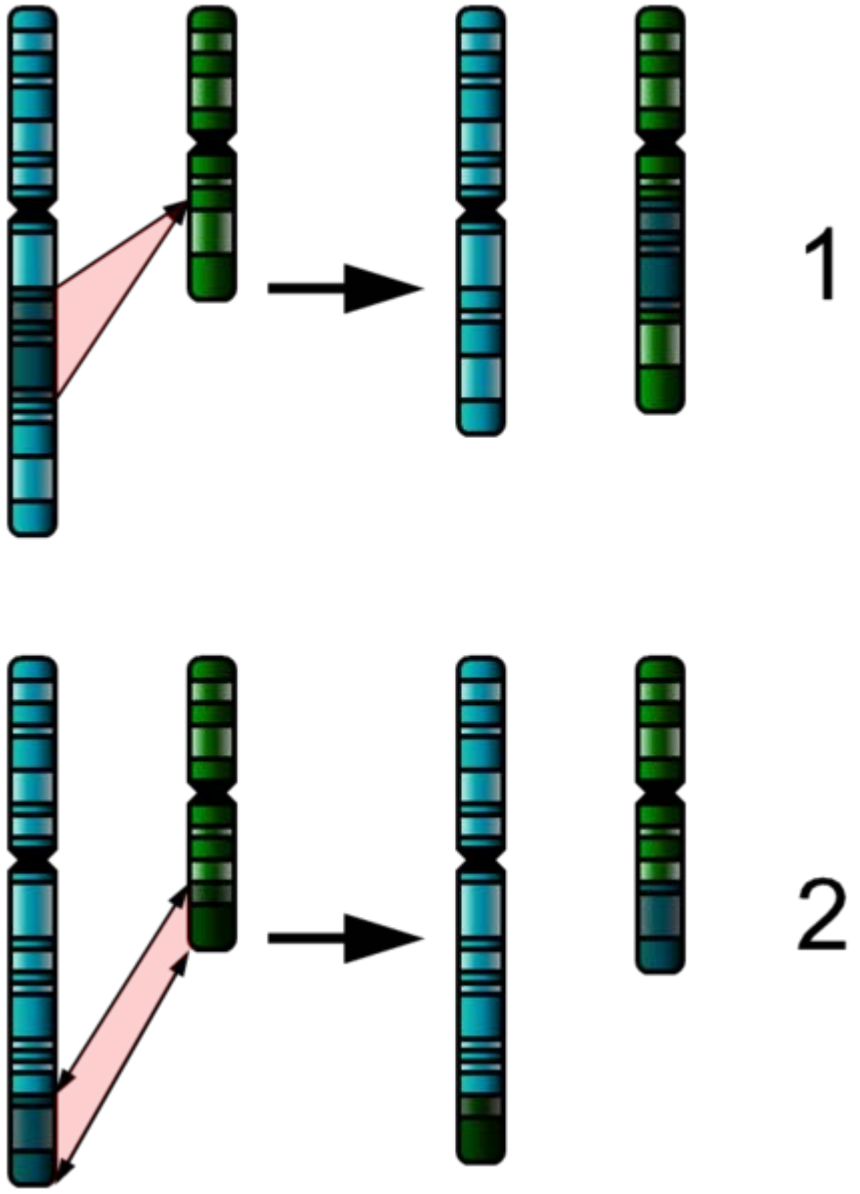
1. Using cells in culture
2. Pretreating cells in a hypotonic solution, which swells them and spreads the chromosomes
3. Arresting mitosis in metaphase by a solution of colchicine
4. Squashing the preparation on the slide forcing the chromosomes into a single plane
5. Cutting up a photomicrograph and arranging the result into an indisputable karyogram.

It took until the mid-1950s for it to become generally accepted that the human karyotype include only 46 chromosomes. Considering the techniques of Winiwarter and Painter, their results were quite remarkable. Chimpanzees (the closest living relatives to modern humans) have 48 chromosomes.

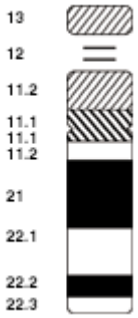
Chromosomal aberrations



The three major single chromosome mutations; deletion (1), duplication (2) and inversion (3)



The two major two-chromosome mutations; insertion (1) and translocation (2)



In Down syndrome, there are three copies of chromosome 21

Chromosomal aberrations are disruptions in the normal chromosomal content of a cell and are a major cause of genetic conditions in humans, such as Down syndrome. Some chromosome abnormalities do not cause disease in carriers, such as translocations, or chromosomal inversions, although they may lead to a higher chance of birthing a child with a chromosome disorder. Abnormal numbers of chromosomes or chromosome sets, aneuploidy, may be lethal or give rise to genetic disorders. Genetic counseling is offered for families that may carry a chromosome rearrangement.

The gain or loss of DNA from chromosomes can lead to a variety of genetic disorders. Human examples include:

- Cri du chat, which is caused by the deletion of part of the short arm of chromosome 5. "Cri du chat" means "cry of the cat" in French, and the condition was so-named because affected babies make high-pitched cries that sound like those of a cat. Affected individuals have wide-set eyes, a small head and jaw, moderate to severe mental health issues, and are very short.
- Down syndrome, usually is caused by an extra copy of chromosome 21 (trisomy 21). Characteristics include decreased muscle tone, stockier build, asymmetrical skull, slanting eyes and mild to moderate developmental disability.
- Edwards syndrome, which is the second-most-common trisomy; Down syndrome is the most common. It is a trisomy of chromosome 18. Symptoms include motor retardation, developmental disability and numerous congenital anomalies causing serious health problems. Ninety percent die in infancy; however, those that live past their first birthday usually are quite healthy thereafter. They have a characteristic clenched hands and overlapping fingers.
- Idic15, abbreviation for Isodicentric 15 on chromosome 15; also called the following names due to various researches, but they all mean the same; IDIC(15), Inverted duplication 15, extra Marker, Inv dup 15, partial tetrasomy 15
- Jacobsen syndrome, also called the terminal 11q deletion disorder. This is a very rare disorder. Those affected have normal intelligence or mild developmental disability, with poor expressive language skills. Most have a bleeding disorder called Paris-Trousseau syndrome.
- Klinefelter's syndrome (XXY). Men with Klinefelter syndrome are usually sterile, and tend to have longer arms and legs and to be taller than their peers. Boys with the syndrome are often shy and quiet, and have a higher incidence of speech delay and dyslexia. During puberty, without testosterone treatment, some of them may develop gynecomastia.
- Patau Syndrome, also called D-Syndrome or trisomy-13. Symptoms are somewhat similar to those of trisomy-18, but they do not have the characteristic hand shape.
- Small supernumerary marker chromosome. This means there is an extra, abnormal chromosome. Features depend on the origin of the extra genetic material. Cat-eye syndrome and isodicentric chromosome 15 syndrome (or Idic15) are both caused by a supernumerary marker chromosome, as is Pallister-Killian syndrome.

- Triple-X syndrome (XXX). XXX girls tend to be tall and thin. They have a higher incidence of dyslexia.
- Turner syndrome (X instead of XX or XY). In Turner syndrome, female sexual characteristics are present but underdeveloped. People with Turner syndrome often have a short stature, low hairline, abnormal eye features and bone development and a "caved-in" appearance to the chest.
- XYY syndrome. XYY boys are usually taller than their siblings. Like XXY boys and XXX girls, they are somewhat more likely to have learning difficulties.
- Wolf-Hirschhorn syndrome, which is caused by partial deletion of the short arm of chromosome 4. It is characterized by severe growth retardation and severe to profound mental health issues.

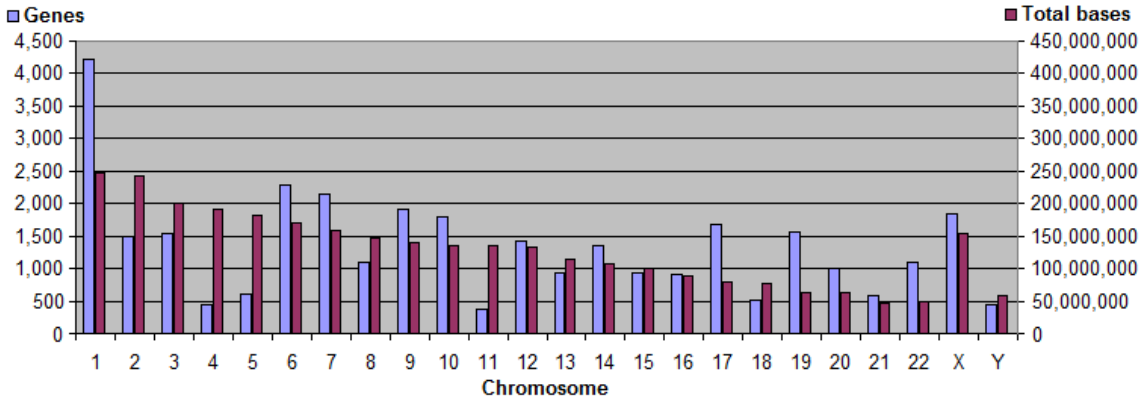
Chromosomal mutations produce changes in whole chromosomes (more than one gene) or in the number of chromosomes present.

- Deletion – loss of part of a chromosome
- Duplication – extra copies of a part of a chromosome
- Inversion – reverse the direction of a part of a chromosome
- Translocation – part of a chromosome breaks off and attaches to another chromosome

Most mutations are neutral – have little or no effect. Chromosomal aberrations are the changes in the structure of chromosomes. It has a great role in evolution. A detailed graphical display of all human chromosomes and the diseases annotated at the correct spot may be found at the Oak Ridge National Laboratory.

Human chromosomes

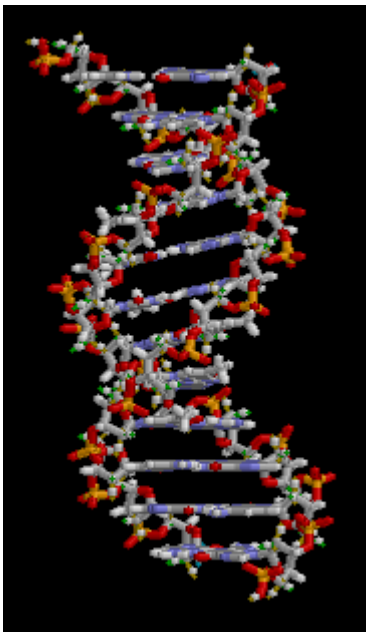
Chromosomes can be divided into two types—autosomes, and sex chromosomes. Certain genetic traits are linked to your sex, and are passed on through the sex chromosomes. The autosomes contain the rest of the genetic hereditary information. All act in the same way during cell division. Human cells have 23 pairs of large linear nuclear chromosomes, (22 pairs of autosomes and one pair of sex chromosomes) giving a total of 46 per cell. In addition to these, human cells have many hundreds of copies of the mitochondrial genome. Sequencing of the human genome has provided a great deal of information about each of the chromosomes. Below is a table compiling statistics for the chromosomes, based on the Sanger Institute's human genome information in the Vertebrate Genome Annotation (VEGA) database. Number of genes is an estimate as it is in part based on gene predictions. Total chromosome length is an estimate as well, based on the estimated size of unsequenced heterochromatin regions.



Chromosome	Genes	Total bases	Sequenced bases
1	4,220	247,199,719	224,999,719
2	1,491	242,751,149	237,712,649
3	1,550	199,446,827	194,704,827
4	446	191,263,063	187,297,063
5	609	180,837,866	177,702,766
6	2,281	170,896,993	167,273,993
7	2,135	158,821,424	154,952,424
8	1,106	146,274,826	142,612,826
9	1,920	140,442,298	120,312,298
10	1,793	135,374,737	131,624,737
11	379	134,452,384	131,130,853
12	1,430	132,289,534	130,303,534
13	924	114,127,980	95,559,980
14	1,347	106,360,585	88,290,585
15	921	100,338,915	81,341,915
16	909	88,822,254	78,884,754
17	1,672	78,654,742	77,800,220
18	519	76,117,153	74,656,155
19	1,555	63,806,651	55,785,651
20	1,008	62,435,965	59,505,254
21	578	46,944,323	34,171,998
22	1,092	49,528,953	34,893,953
X (sex chromosome)	1,846	154,913,754	151,058,754
Y (sex chromosome)	454	57,741,652	25,121,652
Total	**32,185**	**3,079,843,747**	**2,857,698,560**

Chapter- 2

DNA



The structure of part of a DNA double helix

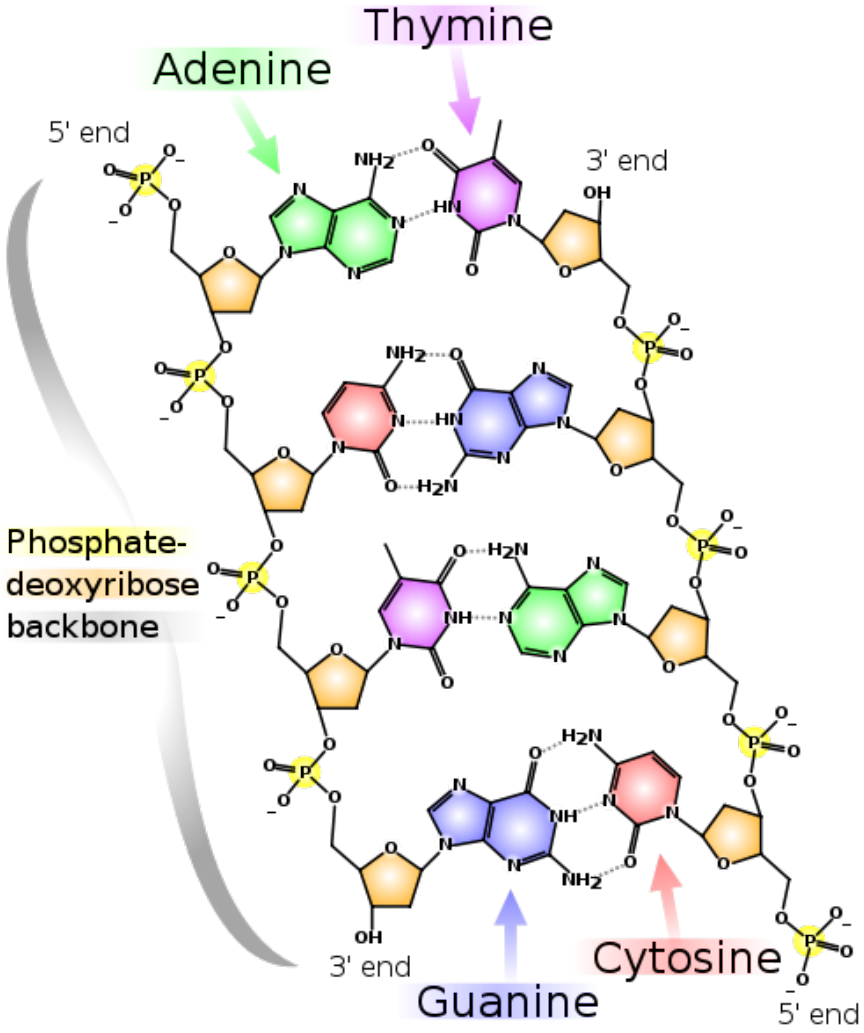
Deoxyribonucleic acid, or **DNA**, is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms (with the exception of RNA viruses). The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set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

DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are therefore anti-parallel. Attached to each sugar is one of four types of molecules called bases. It is the sequence of these four bases along

the backbone that encodes information. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related nucleic acid RNA, in a process called transcription.

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

Properties

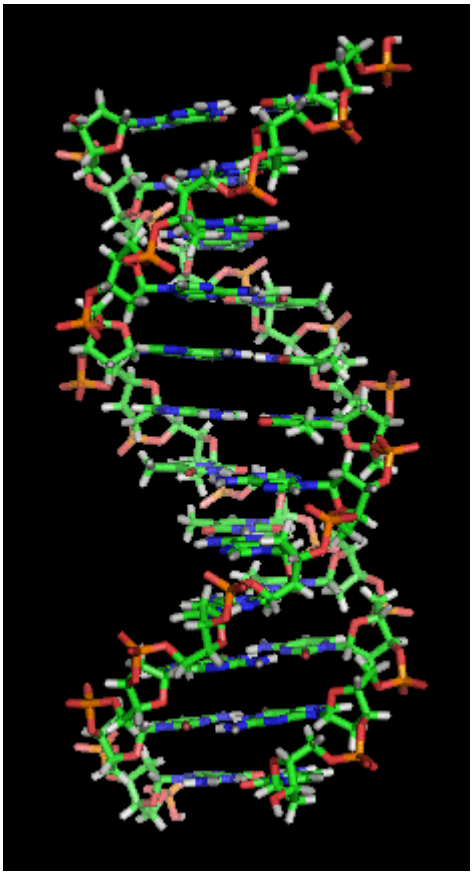


Chemical structure of DNA. Hydrogen bonds shown as dotted lines

DNA is a long polymer made from repeating units called nucleotides. As first discovered by James D. Watson and Francis Crick, the structure of DNA of all species comprises two helical chains each coiled round the same axis, and each with a pitch of 34 Ångströms (3.4 nanometres) and a radius of 10 Ångströms (1.0 nanometres). According to another study, when measured in a particular solution, the DNA chain measured 22 to 26 Ångströms wide (2.2 to 2.6 nanometres), and one nucleotide unit measured 3.3 Å (0.33 nm) long. Although each individual repeating unit is very small, DNA polymers can be very large molecules containing millions of nucleotides. For instance, the largest human chromosome, chromosome number 1, is approximately 220 million base pairs long.

In living organisms, DNA does not usually exist as a single molecule, but instead as a pair of molecules that are held tightly together. These two long strands entwine like vines, in the shape of a double helix. The nucleotide repeats contain both the segment of the backbone of the molecule, which holds the chain together, and a base, which interacts with the other DNA strand in the helix. A base linked to a sugar is called a nucleoside and a base linked to a sugar and one or more phosphate groups is called a nucleotide. If multiple nucleotides are linked together, as in DNA, this polymer is called a polynucleotide.

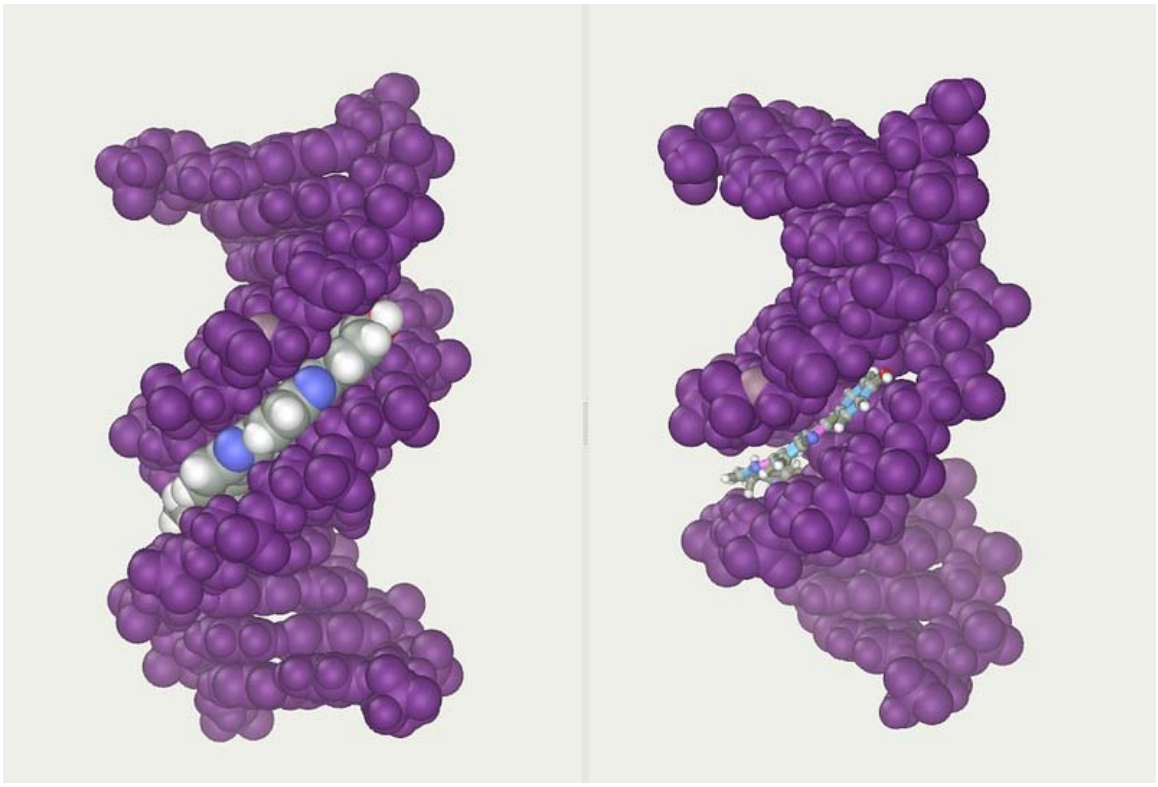
The backbone of the DNA strand is made from alternating phosphate and sugar residues. The sugar in DNA is 2-deoxyribose, which is a pentose (five-carbon) sugar. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These asymmetric bonds mean a strand of DNA has a direction. In a double helix the direction of the nucleotides in one strand is opposite to their direction in the other strand: the strands are *antiparallel*. The asymmetric ends of DNA strands are called the 5′ (*five prime*) and 3′ (*three prime*) ends, with the 5' end having a terminal phosphate group and the 3' end a terminal hydroxyl group. One major difference between DNA and RNA is the sugar, with the 2-deoxyribose in DNA being replaced by the alternative pentose sugar ribose in RNA.



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

The DNA double helix is stabilized by hydrogen bonds between the bases attached to the two strands. The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). These four bases are attached to the sugar/phosphate to form the complete nucleotide, as shown for adenosine monophosphate.

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



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

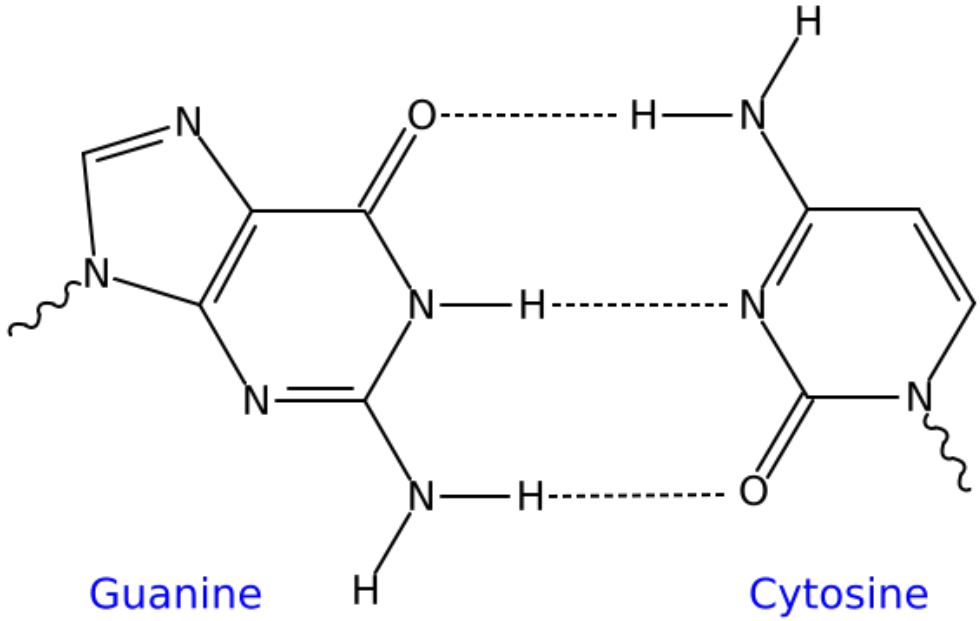
Grooves

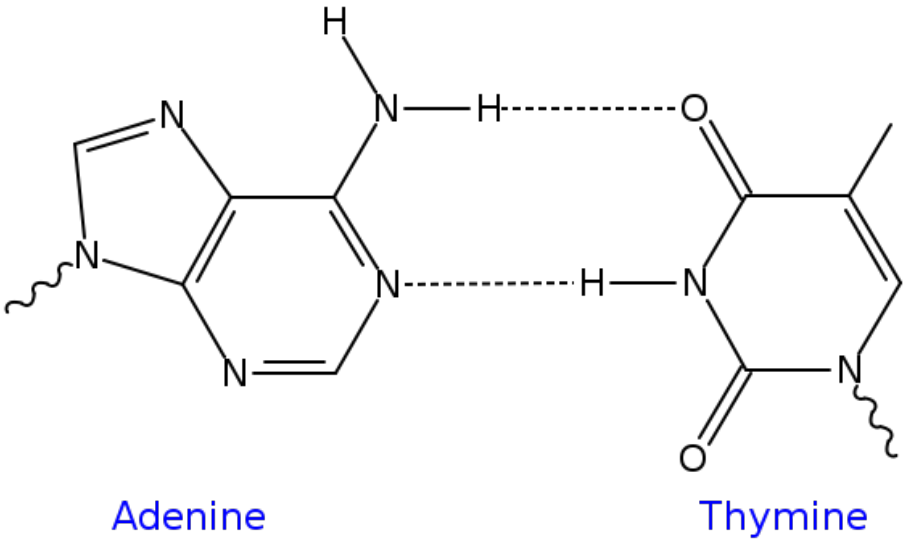
Twin helical strands form the DNA backbone. Another double helix may be found by tracing the spaces, or grooves, between the strands. These voids are adjacent to the base pairs and may provide a binding site. As the strands are not directly opposite each other, the grooves are unequally sized. One groove, the major groove, is 22 Å wide and the other, the minor groove, is 12 Å wide. The narrowness of the minor groove means that the edges of the bases are more accessible in the major groove. As a result, proteins like transcription factors that can bind to specific sequences in double-stranded DNA usually make contacts to the sides of the bases exposed in the major groove. This situation varies in unusual conformations of DNA within the cell *(see below)*, but the major and minor grooves are always named to reflect the differences in size that would be seen if the DNA is twisted back into the ordinary B form.

Base pairing

Each type of base on one strand forms a bond with just one type of base on the other strand. This is called complementary base pairing. Here, purines form hydrogen bonds to pyrimidines, with A bonding only to T, and C bonding only to G. This arrangement of two nucleotides binding together across the double helix is called a base pair. As hydrogen bonds are not covalent, they can be broken and rejoined relatively easily. The

two strands of DNA in a double helix can therefore be pulled apart like a zipper, either by a mechanical force or high temperature. As a result of this complementarity, all the information in the double-stranded sequence of a DNA helix is duplicated on each strand, which is vital in DNA replication. Indeed, this reversible and specific interaction between complementary base pairs is critical for all the functions of DNA in living organisms.





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

The two types of base pairs form different numbers of hydrogen bonds, AT forming two hydrogen bonds, and GC forming three hydrogen bonds. DNA with high GC-content is more stable than DNA with low GC-content, but contrary to popular belief, this is not due to the extra hydrogen bond of a GC base pair but rather the contribution of stacking interactions (hydrogen bonding merely provides specificity of the pairing, not stability).

As a result, it is both the percentage of GC base pairs and the overall length of a DNA double helix that determine the strength of the association between the two strands of DNA. Long DNA helices with a high GC content have stronger-interacting strands, while short helices with high AT content have weaker-interacting strands. In biology, parts of the DNA double helix that need to separate easily, such as the TATAAT Pribnow box in some promoters, tend to have a high AT content, making the strands easier to pull apart.

In the laboratory, the strength of this interaction can be measured by finding the temperature required to break the hydrogen bonds, their melting temperature (also called T_m value). When all the base pairs in a DNA double helix melt, the strands separate and exist in solution as two entirely independent molecules. These single-stranded DNA molecules (*ssDNA*) have no single common shape, but some conformations are more stable than others.

Sense and antisense

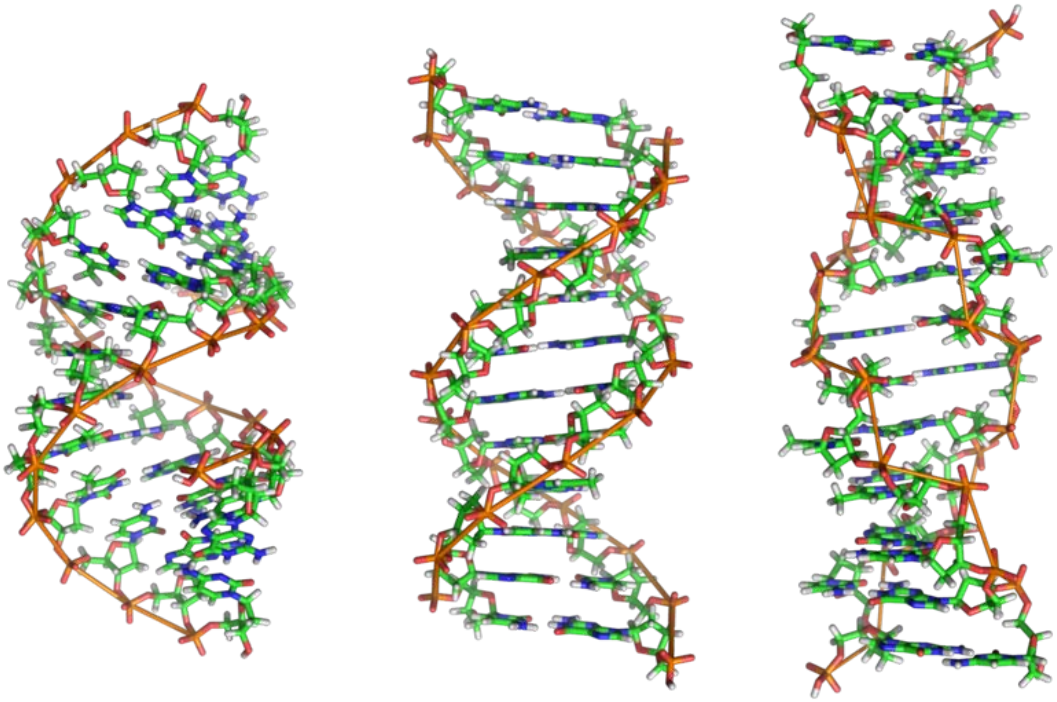
A DNA sequence is called "sense" if its sequence is the same as that of a messenger RNA copy that is translated into protein. The sequence on the opposite strand is called the "antisense" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands contain both sense and antisense sequences). In both prokaryotes and eukaryotes, antisense RNA sequences are produced, but the functions of these RNAs are not entirely clear. One proposal is that antisense RNAs are involved in regulating gene expression through RNA-RNA base pairing.

A few DNA sequences in prokaryotes and eukaryotes, and more in plasmids and viruses, blur the distinction between sense and antisense strands by having overlapping genes. In these cases, some DNA sequences do double duty, encoding one protein when read along one strand, and a second protein when read in the opposite direction along the other strand. In bacteria, this overlap may be involved in the regulation of gene transcription, while in viruses, overlapping genes increase the amount of information that can be encoded within the small viral genome.

Supercoiling

DNA can be twisted like a rope in a process called DNA supercoiling. With DNA in its "relaxed" state, a strand usually circles the axis of the double helix once every 10.4 base pairs, but if the DNA is twisted the strands become more tightly or more loosely wound. If the DNA is twisted in the direction of the helix, this is positive supercoiling, and the bases are held more tightly together. If they are twisted in the opposite direction, this is negative supercoiling, and the bases come apart more easily. In nature, most DNA has

slight negative supercoiling that is introduced by enzymes called topoisomerases. These enzymes are also needed to relieve the twisting stresses introduced into DNA strands during processes such as transcription and DNA replication.



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

Alternate DNA structures

DNA exists in many possible conformations that include A-DNA, B-DNA, and Z-DNA forms, although, only B-DNA and Z-DNA have been directly observed in functional organisms. The conformation that DNA adopts depends on the hydration level, DNA sequence, the amount and direction of supercoiling, chemical modifications of the bases, the type and concentration of metal ions, as well as the presence of polyamines in solution.

The first published reports of A-DNA X-ray diffraction patterns— and also B-DNA used analyses based on Patterson transforms that provided only a limited amount of structural information for oriented fibers of DNA. An alternate analysis was then proposed by Wilkins *et al.*, in 1953, for the *in vivo* B-DNA X-ray diffraction/scattering patterns of highly hydrated DNA fibers in terms of squares of Bessel functions. In the same journal, James D. Watson and Francis Crick presented their molecular modeling analysis of the DNA X-ray diffraction patterns to suggest that the structure was a double-helix.

Although the `B-DNA form' is most common under the conditions found in cells, it is not a well-defined conformation but a family of related DNA conformations that occur at the high hydration levels present in living cells. Their corresponding X-ray diffraction and

scattering patterns are characteristic of molecular paracrystals with a significant degree of disorder.

Compared to B-DNA, the A-DNA form is a wider right-handed spiral, with a shallow, wide minor groove and a narrower, deeper major groove. The A form occurs under non-physiological conditions in partially dehydrated samples of DNA, while in the cell it may be produced in hybrid pairings of DNA and RNA strands, as well as in enzyme-DNA complexes. Segments of DNA where the bases have been chemically modified by methylation may undergo a larger change in conformation and adopt the Z form. Here, the strands turn about the helical axis in a left-handed spiral, the opposite of the more common B form. These unusual structures can be recognized by specific Z-DNA binding proteins and may be involved in the regulation of transcription.

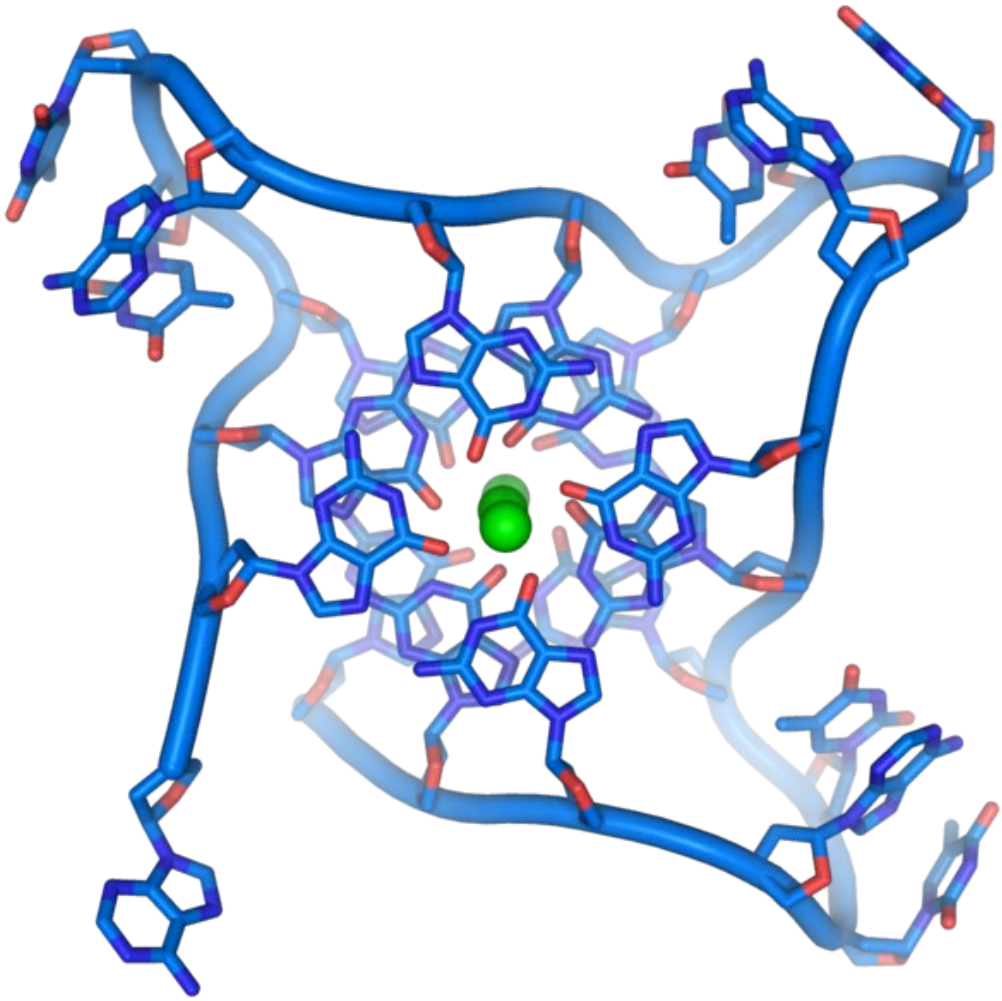
Alternate DNA chemistry

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

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

Quadruplex structures

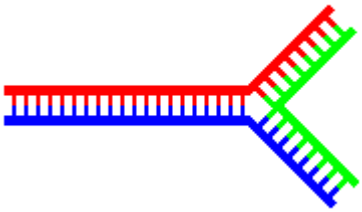
At the ends of the linear chromosomes are specialized regions of DNA called telomeres. The main function of these regions is to allow the cell to replicate chromosome ends using the enzyme telomerase, as the enzymes that normally replicate DNA cannot copy the extreme 3′ ends of chromosomes. These specialized chromosome caps also help protect the DNA ends, and stop the DNA repair systems in the cell from treating them as damage to be corrected. In human cells, telomeres are usually lengths of single-stranded DNA containing several thousand repeats of a simple TTAGGG sequence.



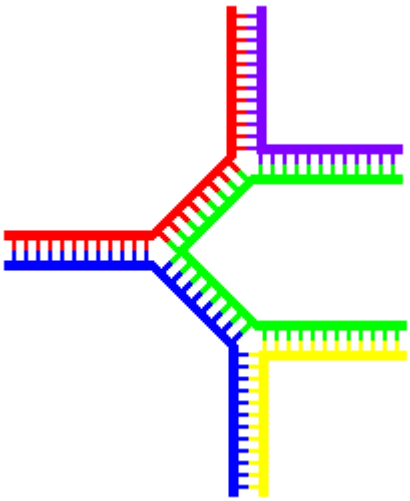
DNA quadruplex formed by telomere repeats. The looped conformation of the DNA backbone is very different from the typical DNA helix.

These guanine-rich sequences may stabilize chromosome ends by forming structures of stacked sets of four-base units, rather than the usual base pairs found in other DNA molecules. Here, four guanine bases form a flat plate and these flat four-base units then stack on top of each other, to form a stable *G-quadruplex* structure. These structures are stabilized by hydrogen bonding between the edges of the bases and chelation of a metal ion in the centre of each four-base unit. Other structures can also be formed, with the central set of four bases coming from either a single strand folded around the bases, or several different parallel strands, each contributing one base to the central structure.

In addition to these stacked structures, telomeres also form large loop structures called telomere loops, or T-loops. Here, the single-stranded DNA curls around in a long circle stabilized by telomere-binding proteins. At the very end of the T-loop, the single-stranded telomere DNA is held onto a region of double-stranded DNA by the telomere strand disrupting the double-helical DNA and base pairing to one of the two strands. This triple-stranded structure is called a displacement loop or D-loop.



Single branch



Multiple branches

Branched DNA can form networks containing multiple branches.

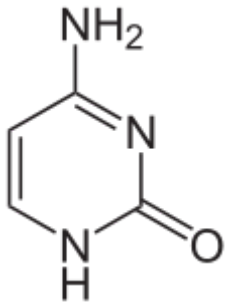
Branched DNA

In DNA fraying occurs when non-complementary regions exist at the end of an otherwise complementary double-strand of DNA. However, branched DNA can occur if a third strand of DNA is introduced and contains adjoining regions able to hybridize with the frayed regions of the pre-existing double-strand. Although the simplest example of branched DNA involves only three strands of DNA, complexes involving additional strands and multiple branches are also possible. Branched DNA can be used in nanotechnology to construct geometric shapes, see the section on uses in technology below.

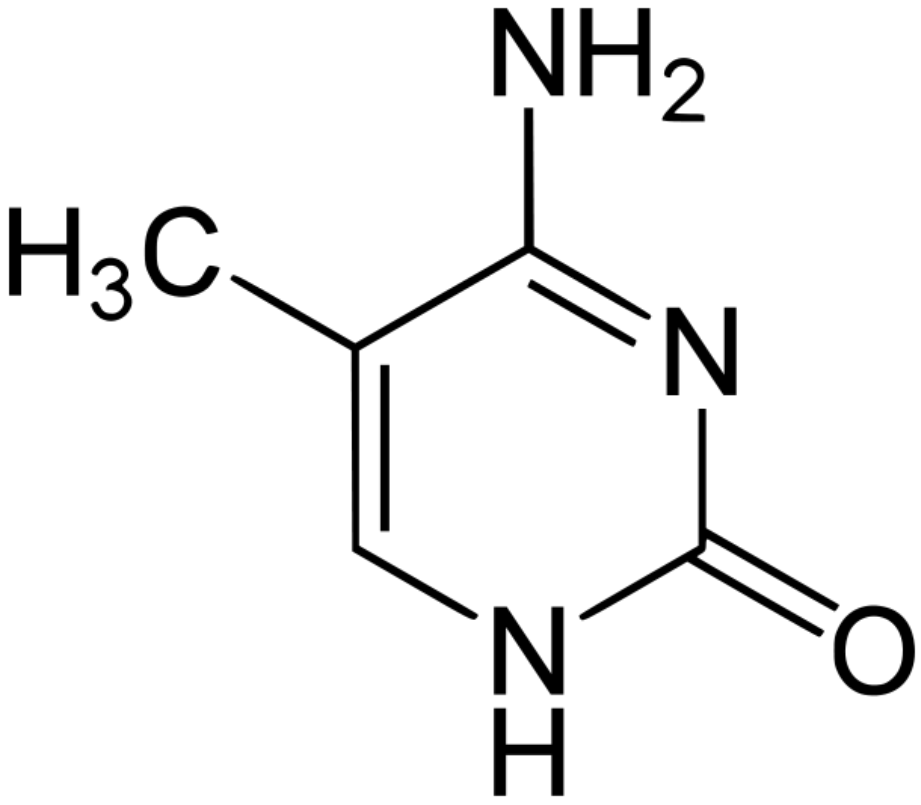
Vibration

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

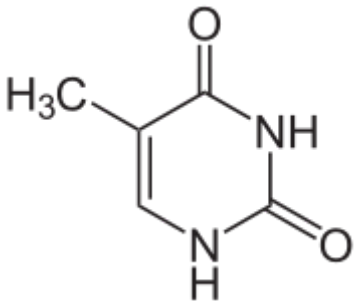
Chemical modifications



cytosine



5-methylcytosine



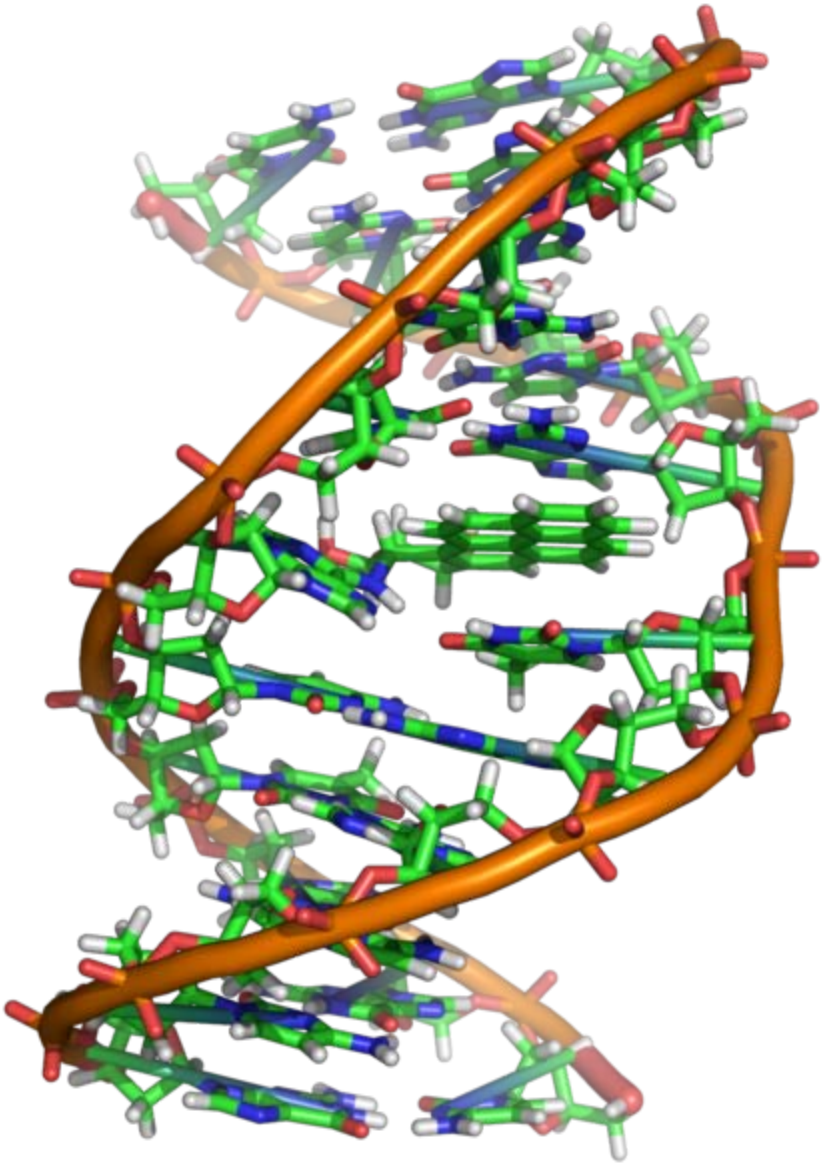
thymine

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

Base modifications

The expression of genes is influenced by how the DNA is packaged in chromosomes, in a structure called chromatin. Base modifications can be involved in packaging, with regions that have low or no gene expression usually containing high levels of methylation of cytosine bases. For example, cytosine methylation, produces 5-methylcytosine, which is important for X-chromosome inactivation. The average level of methylation varies between organisms - the worm *Caenorhabditis elegans* lacks cytosine methylation, while vertebrates have higher levels, with up to 1% of their DNA containing 5-methylcytosine. Despite the importance of 5-methylcytosine, it can deaminate to leave a thymine base, so methylated cytosines are particularly prone to mutations. Other base modifications include adenine methylation in bacteria, the presence of 5-hydroxymethylcytosine in the brain, and the glycosylation of uracil to produce the "J-base" in kinetoplastids.

Damage



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

DNA can be damaged by many sorts of mutagens, which change the DNA sequence. Mutagens include oxidizing agents, alkylating agents and also high-energy electromagnetic radiation such as ultraviolet light and X-rays. The type of DNA damage produced depends on the type of mutagen. For example, UV light can damage DNA by producing thymine dimers, which are cross-links between pyrimidine bases. On the other hand, oxidants such as free radicals or hydrogen peroxide produce multiple forms of damage, including base modifications, particularly of guanosine, and double-strand breaks. A typical human cell contains about 150,000 bases that have suffered oxidative damage. Of these oxidative lesions, the most dangerous are double-strand breaks, as these

are difficult to repair and can produce point mutations, insertions and deletions from the DNA sequence, as well as chromosomal translocations.

Many mutagens fit into the space between two adjacent base pairs, this is called *intercalation*. Most intercalators are aromatic and planar molecules; examples include ethidium bromide, daunomycin, and doxorubicin. In order for an intercalator to fit between base pairs, the bases must separate, distorting the DNA strands by unwinding of the double helix. This inhibits both transcription and DNA replication, causing toxicity and mutations. As a result, DNA intercalators are often carcinogens, and benzo[*a*]pyrene diol epoxide, acridines, aflatoxin and ethidium bromide are well-known examples. Nevertheless, due to their ability to inhibit DNA transcription and replication, other similar toxins are also used in chemotherapy to inhibit rapidly growing cancer cells.

Biological functions

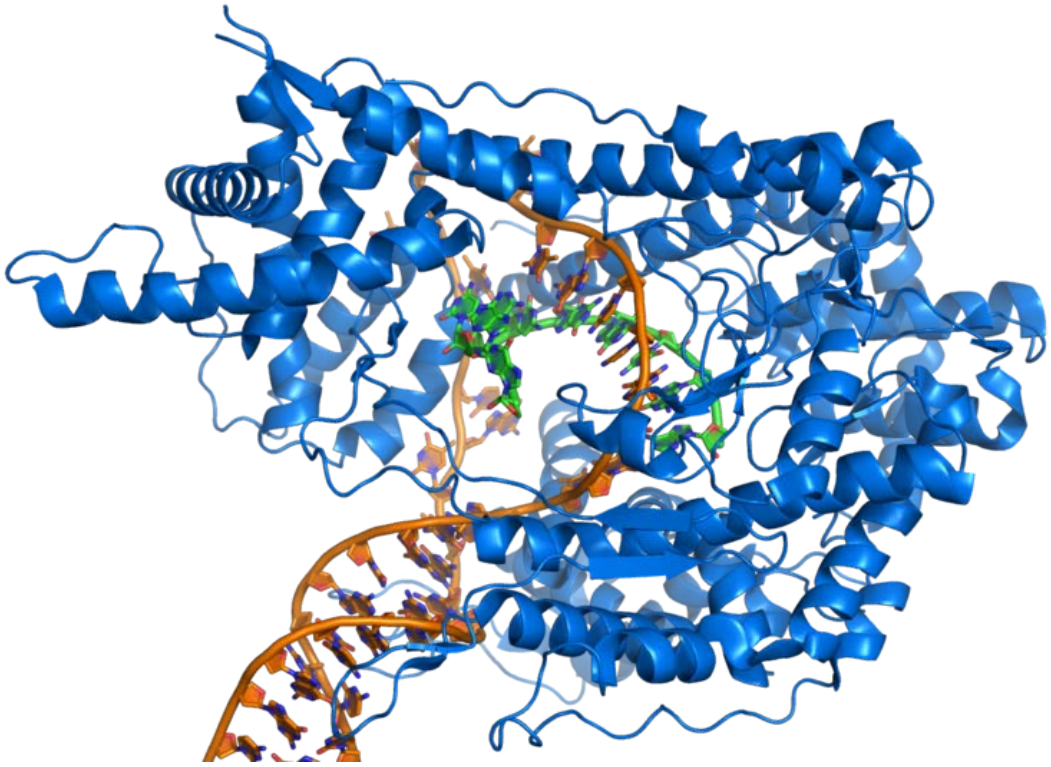
DNA usually occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The set of chromosomes in a cell makes up its genome; the human genome has approximately 3 billion base pairs of DNA arranged into 46 chromosomes. The information carried by DNA is held in the sequence of pieces of DNA called genes. Transmission of genetic information in genes is achieved via complementary base pairing. For example, in transcription, when a cell uses the information in a gene, the DNA sequence is copied into a complementary RNA sequence through the attraction between the DNA and the correct RNA nucleotides. Usually, this RNA copy is then used to make a matching protein sequence in a process called translation, which depends on the same interaction between RNA nucleotides. In alternative fashion, a cell may simply copy its genetic information in a process called DNA replication. The details of these functions are covered in other articles; here we focus on the interactions between DNA and other molecules that mediate the function of the genome.

Genes and genomes

Genomic DNA is tightly and orderly packed in the process called DNA condensation to fit the small available volumes of the cell. In eukaryotes, DNA is located in the cell nucleus, as well as small amounts in mitochondria and chloroplasts. In prokaryotes, the DNA is held within an irregularly shaped body in the cytoplasm called the nucleoid. The genetic information in a genome is held within genes, and the complete set of this information in an organism is called its genotype. A gene is a unit of heredity and is a region of DNA that influences a particular characteristic in an organism. Genes contain an open reading frame that can be transcribed, as well as regulatory sequences such as promoters and enhancers, which control the transcription of the open reading frame.

In many species, only a small fraction of the total sequence of the genome encodes protein. For example, only about 1.5% of the human genome consists of protein-coding exons, with over 50% of human DNA consisting of non-coding repetitive sequences. The reasons for the presence of so much noncoding DNA in eukaryotic genomes and the extraordinary differences in genome size, or *C-value*, among species represent a long-

standing puzzle known as the "C-value enigma". However, DNA sequences that do not code protein may still encode functional non-coding RNA molecules, which are involved in the regulation of gene expression.



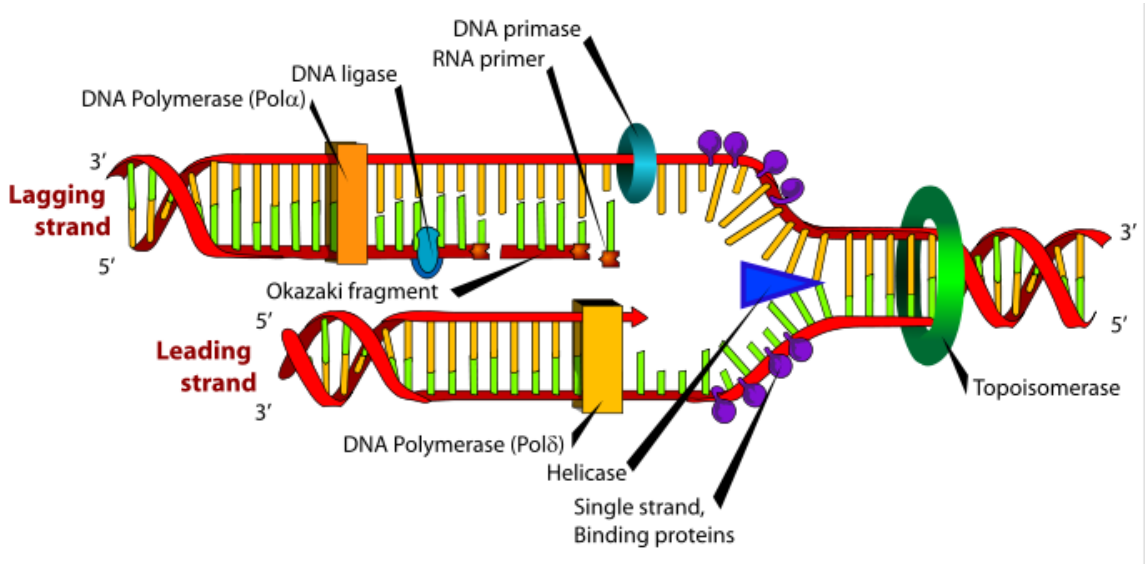
T7 RNA polymerase (blue) producing a mRNA (green) from a DNA template (orange).

Some noncoding DNA sequences play structural roles in chromosomes. Telomeres and centromeres typically contain few genes, but are important for the function and stability of chromosomes. An abundant form of noncoding DNA in humans are pseudogenes, which are copies of genes that have been disabled by mutation. These sequences are usually just molecular fossils, although they can occasionally serve as raw genetic material for the creation of new genes through the process of gene duplication and divergence.

Transcription and translation

A gene is a sequence of DNA that contains genetic information and can influence the phenotype of an organism. Within a gene, the sequence of bases along a DNA strand defines a messenger RNA sequence, which then defines one or more protein sequences. The relationship between the nucleotide sequences of genes and the amino-acid sequences of proteins is determined by the rules of translation, known collectively as the genetic code. The genetic code consists of three-letter 'words' called *codons* formed from a sequence of three nucleotides (e.g. ACT, CAG, TTT).

In transcription, the codons of a gene are copied into messenger RNA by RNA polymerase. This RNA copy is then decoded by a ribosome that reads the RNA sequence by base-pairing the messenger RNA to transfer RNA, which carries amino acids. Since there are 4 bases in 3-letter combinations, there are 64 possible codons (4^3 combinations). These encode the twenty standard amino acids, giving most amino acids more than one possible codon. There are also three 'stop' or 'nonsense' codons signifying the end of the coding region; these are the TAA, TGA and TAG codons.



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

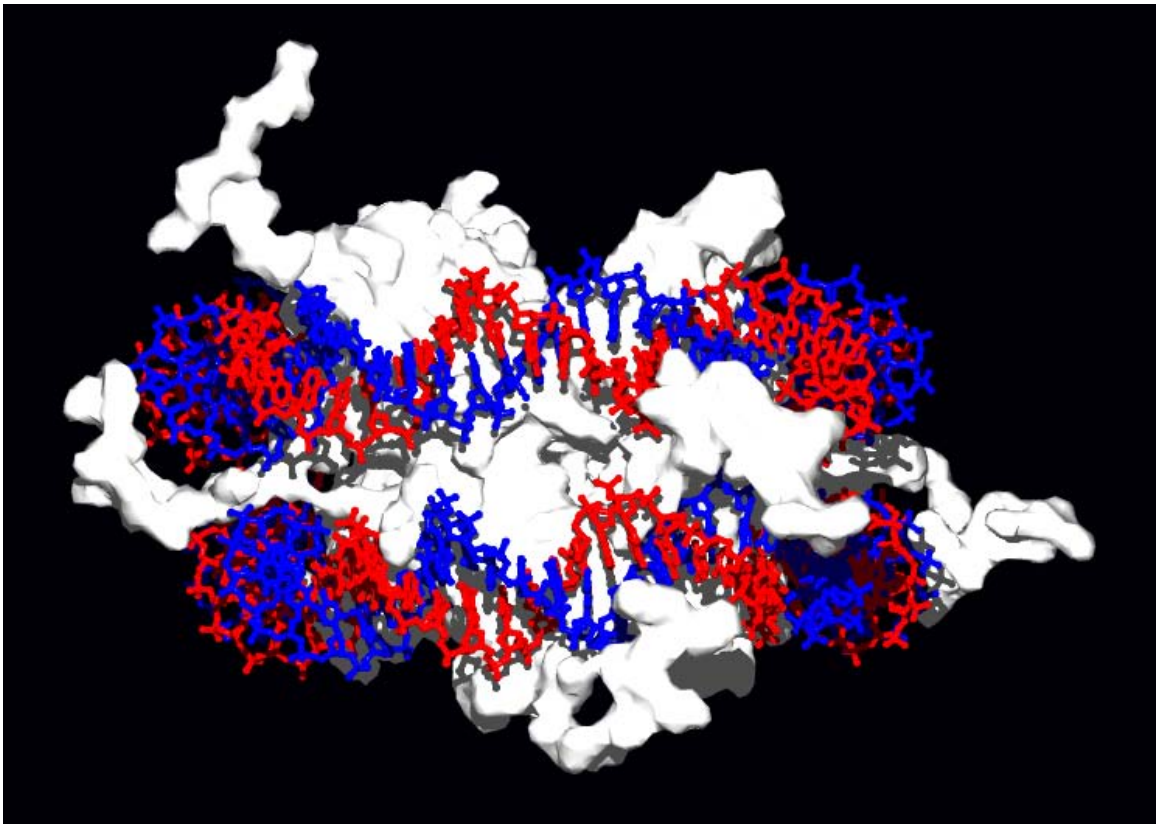
Replication

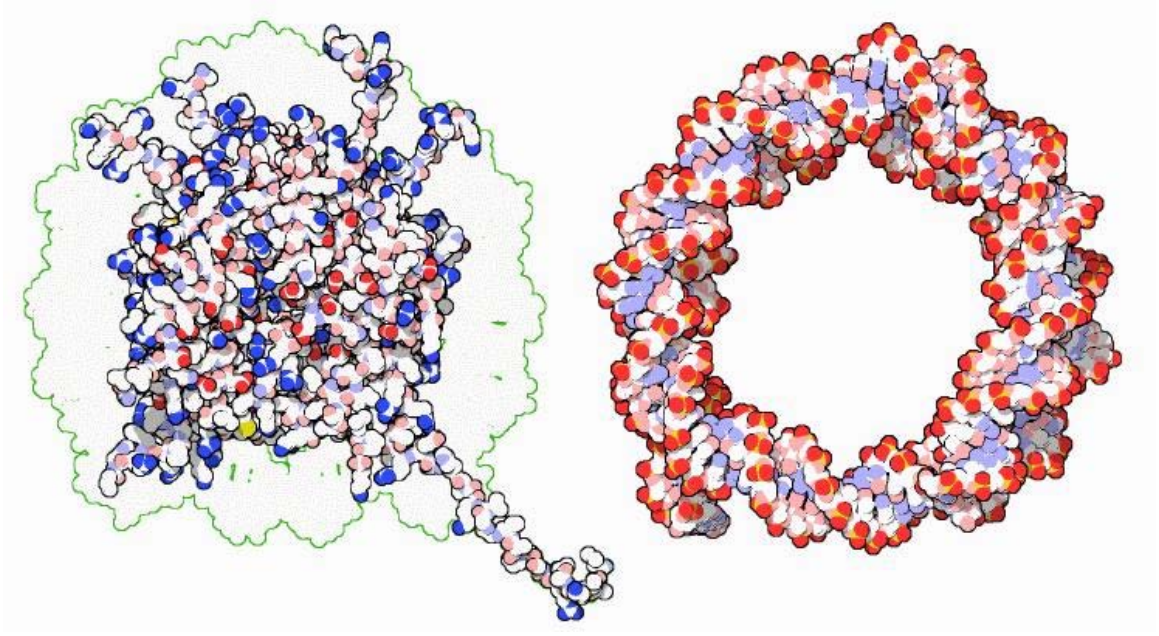
Cell division is essential for an organism to grow, but, when a cell divides, it must replicate the DNA in its genome so that the two daughter cells have the same genetic information as their parent. The double-stranded structure of DNA provides a simple mechanism for DNA replication. Here, the two strands are separated and then each strand's complementary DNA sequence is recreated by an enzyme called DNA polymerase. This enzyme makes the complementary strand by finding the correct base through complementary base pairing, and bonding it onto the original strand. As DNA polymerases can only extend a DNA strand in a 5′ to 3′ direction, different mechanisms are used to copy the antiparallel strands of the double helix. In this way, the base on the old strand dictates which base appears on the new strand, and the cell ends up with a perfect copy of its DNA.

Interactions with proteins

All the functions of DNA depend on interactions with proteins. These protein interactions can be non-specific, or the protein can bind specifically to a single DNA sequence. Enzymes can also bind to DNA and of these, the polymerases that copy the DNA base sequence in transcription and DNA replication are particularly important.

DNA-binding proteins

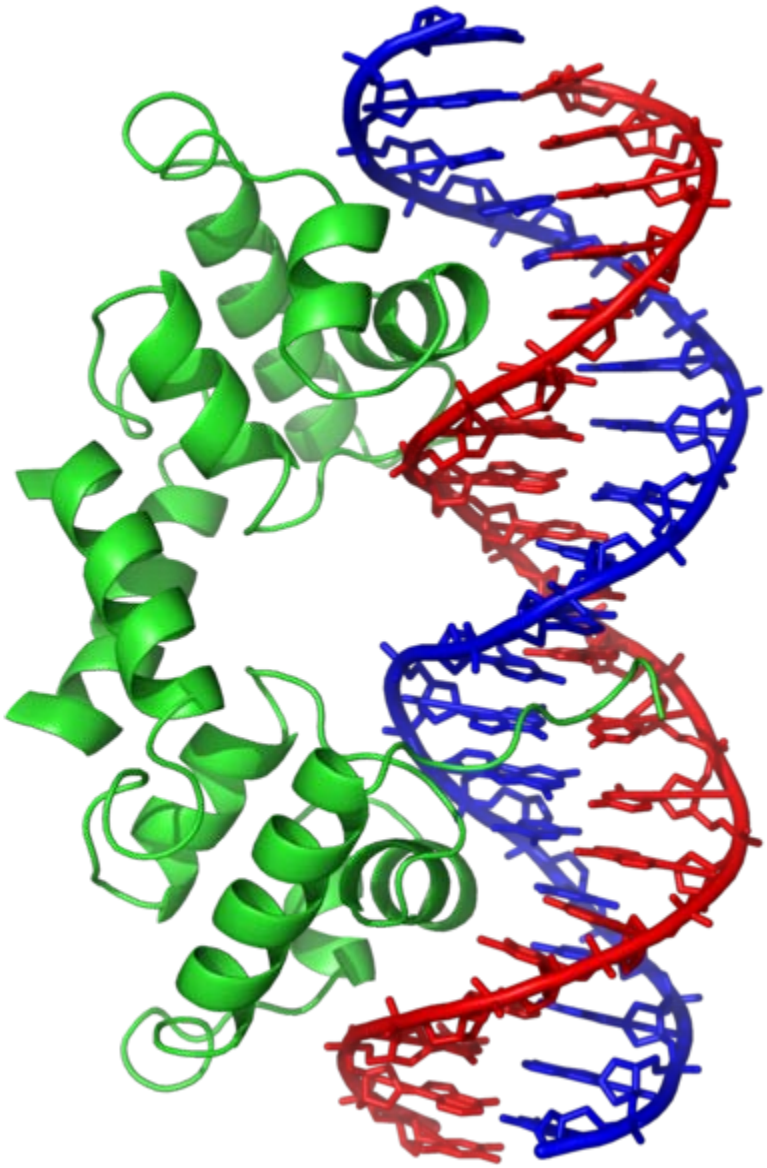




Interaction of DNA with histones (shown in white, top). These proteins' basic amino acids (below left, blue) bind to the acidic phosphate groups on DNA (below right, red).

Structural proteins that bind DNA are well-understood examples of non-specific DNA-protein interactions. Within chromosomes, DNA is held in complexes with structural proteins. These proteins organize the DNA into a compact structure called chromatin. In eukaryotes this structure involves DNA binding to a complex of small basic proteins called histones, while in prokaryotes multiple types of proteins are involved. The histones form a disk-shaped complex called a nucleosome, which contains two complete turns of double-stranded DNA wrapped around its surface. These non-specific interactions are formed through basic residues in the histones making ionic bonds to the acidic sugar-phosphate backbone of the DNA, and are therefore largely independent of the base sequence. Chemical modifications of these basic amino acid residues include methylation, phosphorylation and acetylation. These chemical changes alter the strength of the interaction between the DNA and the histones, making the DNA more or less accessible to transcription factors and changing the rate of transcription. Other non-specific DNA-binding proteins in chromatin include the high-mobility group proteins, which bind to bent or distorted DNA. These proteins are important in bending arrays of nucleosomes and arranging them into the larger structures that make up chromosomes.

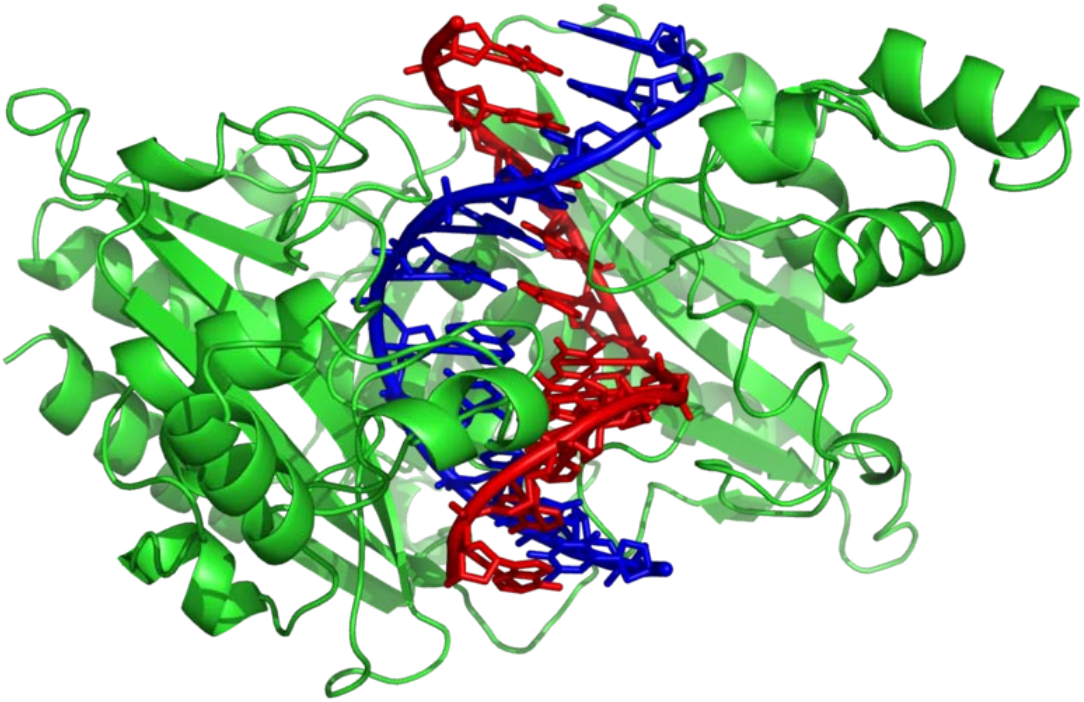
A distinct group of DNA-binding proteins are the DNA-binding proteins that specifically bind single-stranded DNA. In humans, replication protein A is the best-understood member of this family and is used in processes where the double helix is separated, including DNA replication, recombination and DNA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



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

In contrast, other proteins have evolved to bind to particular DNA sequences. The most intensively studied of these are the various transcription factors, which are proteins that regulate transcription. Each transcription factor binds to one particular set of DNA sequences and activates or inhibits the transcription of genes that have these sequences close to their promoters. The transcription factors do this in two ways. Firstly, they can bind the RNA polymerase responsible for transcription, either directly or through other mediator proteins; this locates the polymerase at the promoter and allows it to begin transcription. Alternatively, transcription factors can bind enzymes that modify the histones at the promoter; this will change the accessibility of the DNA template to the polymerase.

As these DNA targets can occur throughout an organism's genome, changes in the activity of one type of transcription factor can affect thousands of genes. Consequently, these proteins are often the targets of the signal transduction processes that control responses to environmental changes or cellular differentiation and development. The specificity of these transcription factors' interactions with DNA come from the proteins making multiple contacts to the edges of the DNA bases, allowing them to "read" the DNA sequence. Most of these base-interactions are made in the major groove, where the bases are most accessible.



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

DNA-modifying enzymes

Nucleases and ligases

Nucleases are enzymes that cut DNA strands by catalyzing the hydrolysis of the phosphodiester bonds. Nucleases that hydrolyse nucleotides from the ends of DNA strands are called exonucleases, while endonucleases cut within strands. The most frequently used nucleases in molecular biology are the restriction endonucleases, which cut DNA at specific sequences. For instance, the EcoRV enzyme shown to the left recognizes the 6-base sequence 5′-GAT|ATC-3′ and makes a cut at the vertical line. In nature, these enzymes protect bacteria against phage infection by digesting the phage DNA when it enters the bacterial cell, acting as part of the restriction modification

system. In technology, these sequence-specific nucleases are used in molecular cloning and DNA fingerprinting.

Enzymes called DNA ligases can rejoin cut or broken DNA strands. Ligases are particularly important in lagging strand DNA replication, as they join together the short segments of DNA produced at the replication fork into a complete copy of the DNA template. They are also used in DNA repair and genetic recombination.

Topoisomerases and helicases

Topoisomerases are enzymes with both nuclease and ligase activity. These proteins change the amount of supercoiling in DNA. Some of these enzymes work by cutting the DNA helix and allowing one section to rotate, thereby reducing its level of supercoiling; the enzyme then seals the DNA break. Other types of these enzymes are capable of cutting one DNA helix and then passing a second strand of DNA through this break, before rejoining the helix. Topoisomerases are required for many processes involving DNA, such as DNA replication and transcription.

Helicases are proteins that are a type of molecular motor. They use the chemical energy in nucleoside triphosphates, predominantly ATP, to break hydrogen bonds between bases and unwind the DNA double helix into single strands. These enzymes are essential for most processes where enzymes need to access the DNA bases.

Polymerases

Polymerases are enzymes that synthesize polynucleotide chains from nucleoside triphosphates. The sequence of their products are copies of existing polynucleotide chains - which are called *templates*. These enzymes function by adding nucleotides onto the 3′ hydroxyl group of the previous nucleotide in a DNA strand. As a consequence, all polymerases work in a 5′ to 3′ direction. In the active site of these enzymes, the incoming nucleoside triphosphate base-pairs to the template: this allows polymerases to accurately synthesize the complementary strand of their template. Polymerases are classified according to the type of template that they use.

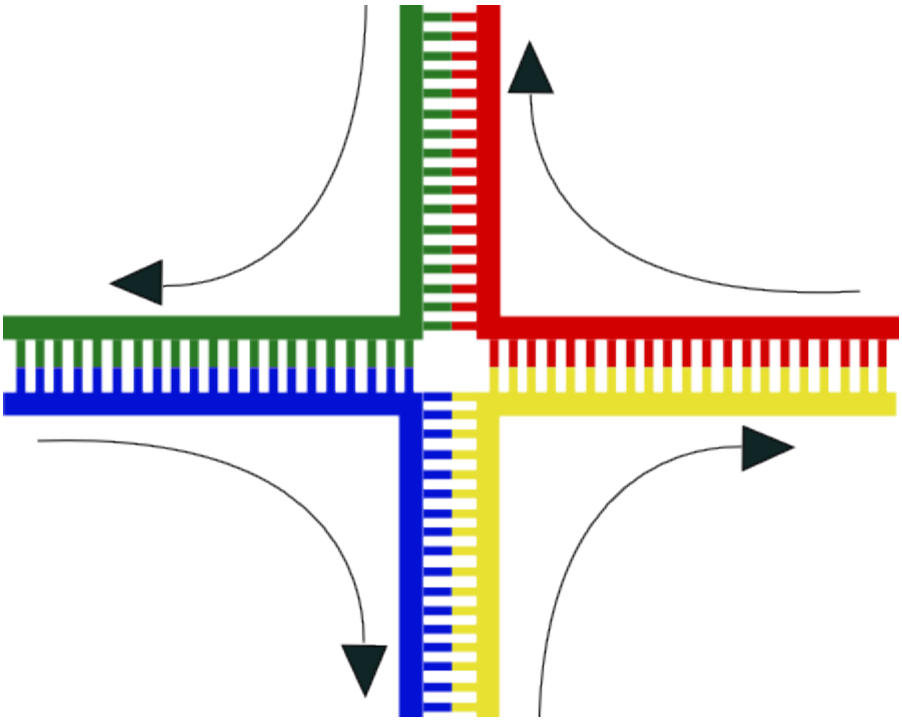
In DNA replication, a DNA-dependent DNA polymerase makes a copy of a DNA sequence. Accuracy is vital in this process, so many of these polymerases have a proofreading activity. Here, the polymerase recognizes the occasional mistakes in the synthesis reaction by the lack of base pairing between the mismatched nucleotides. If a mismatch is detected, a 3′ to 5′ exonuclease activity is activated and the incorrect base removed. In most organisms, DNA polymerases function in a large complex called the replisome that contains multiple accessory subunits, such as the DNA clamp or helicases.

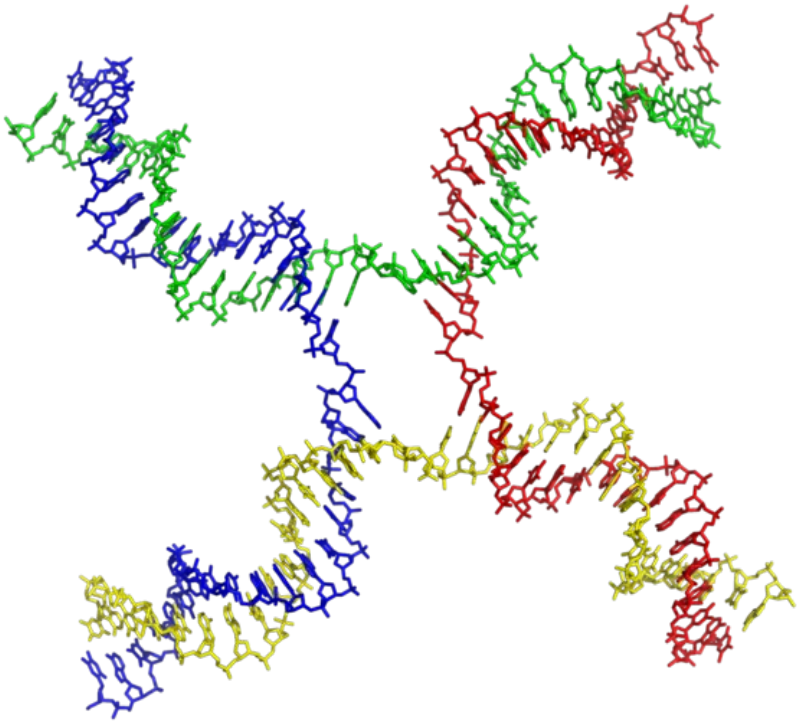
RNA-dependent DNA polymerases are a specialized class of polymerases that copy the sequence of an RNA strand into DNA. They include reverse transcriptase, which is a viral enzyme involved in the infection of cells by retroviruses, and telomerase, which is

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

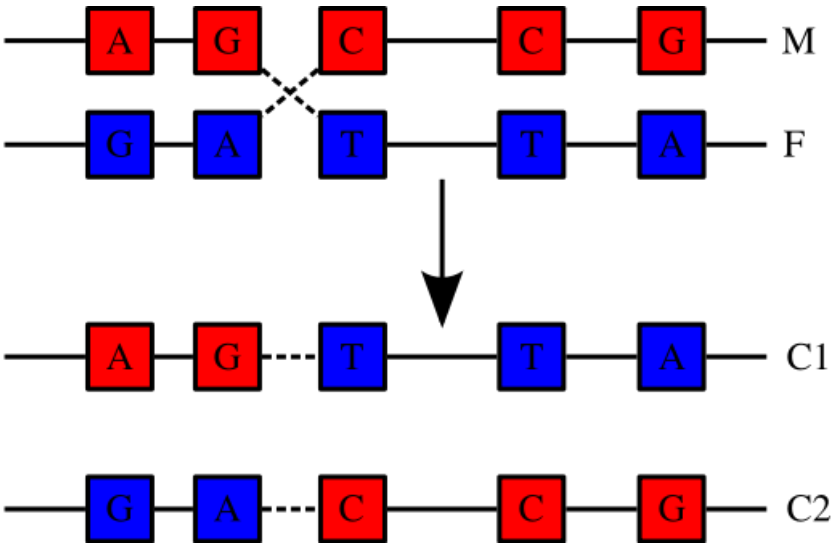
Transcription is carried out by a DNA-dependent RNA polymerase that copies the sequence of a DNA strand into RNA. To begin transcribing a gene, the RNA polymerase binds to a sequence of DNA called a promoter and separates the DNA strands. It then copies the gene sequence into a messenger RNA transcript until it reaches a region of DNA called the terminator, where it halts and detaches from the DNA. As with human DNA-dependent DNA polymerases, RNA polymerase II, the enzyme that transcribes most of the genes in the human genome, operates as part of a large protein complex with multiple regulatory and accessory subunits.

Genetic recombination





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



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

A DNA helix usually does not interact with other segments of DNA, and in human cells the different chromosomes even occupy separate areas in the nucleus called "chromosome territories". This physical separation of different chromosomes is important for the ability of DNA to function as a stable repository for information, as one of the few times chromosomes interact is during chromosomal crossover when they recombine. Chromosomal crossover is when two DNA helices break, swap a section and then rejoin.

Recombination allows chromosomes to exchange genetic information and produces new combinations of genes, which increases the efficiency of natural selection and can be important in the rapid evolution of new proteins. Genetic recombination can also be involved in DNA repair, particularly in the cell's response to double-strand breaks.

The most common form of chromosomal crossover is homologous recombination, where the two chromosomes involved share very similar sequences. Non-homologous recombination can be damaging to cells, as it can produce chromosomal translocations and genetic abnormalities. The recombination reaction is catalyzed by enzymes known as recombinases, such as RAD51. The first step in recombination is a double-stranded break either caused by an endonuclease or damage to the DNA. A series of steps catalyzed in part by the recombinase then leads to joining of the two helices by at least one Holliday junction, in which a segment of a single strand in each helix is annealed to the complementary strand in the other helix. The Holliday junction is a tetrahedral junction structure that can be moved along the pair of chromosomes, swapping one strand for another. The recombination reaction is then halted by cleavage of the junction and re-ligation of the released DNA.

Evolution

DNA contains the genetic information that allows all modern living things to function, grow and reproduce. However, it is unclear how long in the 4-billion-year history of life DNA has performed this function, as it has been proposed that the earliest forms of life may have used RNA as their genetic material. RNA may have acted as the central part of early cell metabolism as it can both transmit genetic information and carry out catalysis as part of ribozymes. This ancient RNA world where nucleic acid would have been used for both catalysis and genetics may have influenced the evolution of the current genetic code based on four nucleotide bases. This would occur, since the number of different bases in such an organism is a trade-off between a small number of bases increasing replication accuracy and a large number of bases increasing the catalytic efficiency of ribozymes.

However, there is no direct evidence of ancient genetic systems, as recovery of DNA from most fossils is impossible. This is because DNA will survive in the environment for less than one million years and slowly degrades into short fragments in solution. Claims for older DNA have been made, most notably a report of the isolation of a viable bacterium from a salt crystal 250 million years old, but these claims are controversial.

Uses in technology

Genetic engineering

Methods have been developed to purify DNA from organisms, such as phenol-chloroform extraction, and to manipulate it in the laboratory, such as restriction digests and the polymerase chain reaction. Modern biology and biochemistry make intensive use of these techniques in recombinant DNA technology. Recombinant DNA is a man-made DNA sequence that has been assembled from other DNA sequences. They can be transformed into organisms in the form of plasmids or in the appropriate format, by using a viral vector. The genetically modified organisms produced can be used to produce products such as recombinant proteins, used in medical research, or be grown in agriculture.

Forensics

Forensic scientists can use DNA in blood, semen, skin, saliva or hair found at a crime scene to identify a matching DNA of an individual, such as a perpetrator. This process is formally termed DNA profiling, but may also be called "genetic fingerprinting". In DNA profiling, the lengths of variable sections of repetitive DNA, such as short tandem repeats and minisatellites, are compared between people. This method is usually an extremely reliable technique for identifying a matching DNA. However, identification can be complicated if the scene is contaminated with DNA from several people. DNA profiling was developed in 1984 by British geneticist Sir Alec Jeffreys, and first used in forensic science to convict Colin Pitchfork in the 1988 Enderby murders case.

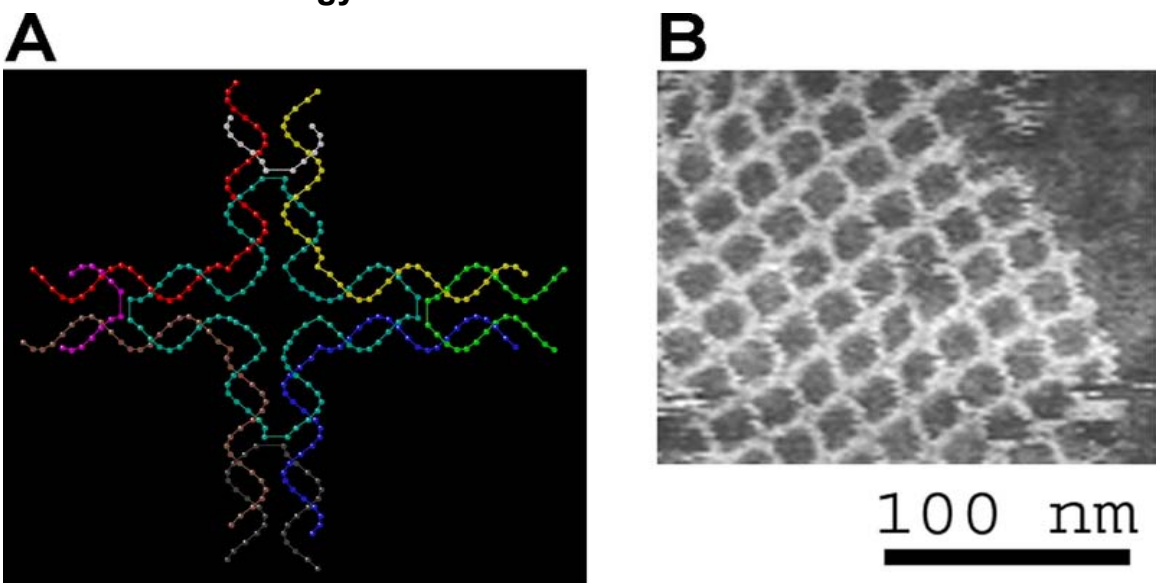
People convicted of certain types of crimes may be required to provide a sample of DNA for a database. This has helped investigators solve old cases where only a DNA sample was obtained from the scene. DNA profiling can also be used to identify victims of mass casualty incidents. On the other hand, many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

Bioinformatics

Bioinformatics involves the manipulation, searching, and data mining of biological data, and this includes DNA sequence data. The development of techniques to store and search DNA sequences have led to widely applied advances in computer science, especially string searching algorithms, machine learning and database theory. String searching or matching algorithms, which find an occurrence of a sequence of letters inside a larger sequence of letters, were developed to search for specific sequences of nucleotides. The DNA sequenced may be aligned with other DNA sequences to identify homologous sequences and locate the specific mutations that make them distinct. These techniques, especially multiple sequence alignment, are used in studying phylogenetic relationships and protein function. Data sets representing entire genomes' worth of DNA sequences, such as those produced by the Human Genome Project, are difficult to use without the

annotations that identify the locations of genes and regulatory elements on each chromosome. Regions of DNA sequence that have the characteristic patterns associated with protein- or RNA-coding genes can be identified by gene finding algorithms, which allow researchers to predict the presence of particular gene products and their possible functions in an organism even before they have been isolated experimentally. Entire genomes may also be compared which can shed light on the evolutionary history of particular organism and permit the examination of complex evolutionary events.

DNA nanotechnology



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

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

History and anthropology

Because DNA collects mutations over time, which are then inherited, it contains historical information, and, by comparing DNA sequences, geneticists can infer the evolutionary history of organisms, their phylogeny. This field of phylogenetics is a powerful tool in evolutionary biology. If DNA sequences within a species are compared,

population geneticists can learn the history of particular populations. This can be used in studies ranging from ecological genetics to anthropology; For example, DNA evidence is being used to try to identify the Ten Lost Tribes of Israel.

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

History of DNA research

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

DNA was first isolated by the Swiss physician Friedrich Miescher who, in 1869, discovered a microscopic substance in the pus of discarded surgical bandages. As it resided in the nuclei of cells, he called it "nuclein". In 1919, Phoebus Levene identified the base, sugar and phosphate nucleotide unit. Levene suggested that DNA consisted of a string of nucleotide units linked together through the phosphate groups. However, Levene thought the chain was short and the bases repeated in a fixed order. In 1937 William

Astbury produced the first X-ray diffraction patterns that showed that DNA had a regular structure.

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

In 1928, Frederick Griffith discovered that traits of the "smooth" form of the *Pneumococcus* could be transferred to the "rough" form of the same bacteria by mixing killed "smooth" bacteria with the live "rough" form. This system provided the first clear suggestion that DNA carries genetic information—the Avery–MacLeod–McCarty experiment—when Oswald Avery, along with coworkers Colin MacLeod and Maclyn McCarty, identified DNA as the transforming principle in 1943. DNA's role in heredity was confirmed in 1952, when Alfred Hershey and Martha Chase in the Hershey–Chase experiment showed that DNA is the genetic material of the T2 phage.

In 1953, James D. Watson and Francis Crick suggested what is now accepted as the first correct double-helix model of DNA structure in the journal *Nature*. Their double-helix, molecular model of DNA was then based on a single X-ray diffraction image (labeled as "Photo 51") taken by Rosalind Franklin and Raymond Gosling in May 1952, as well as

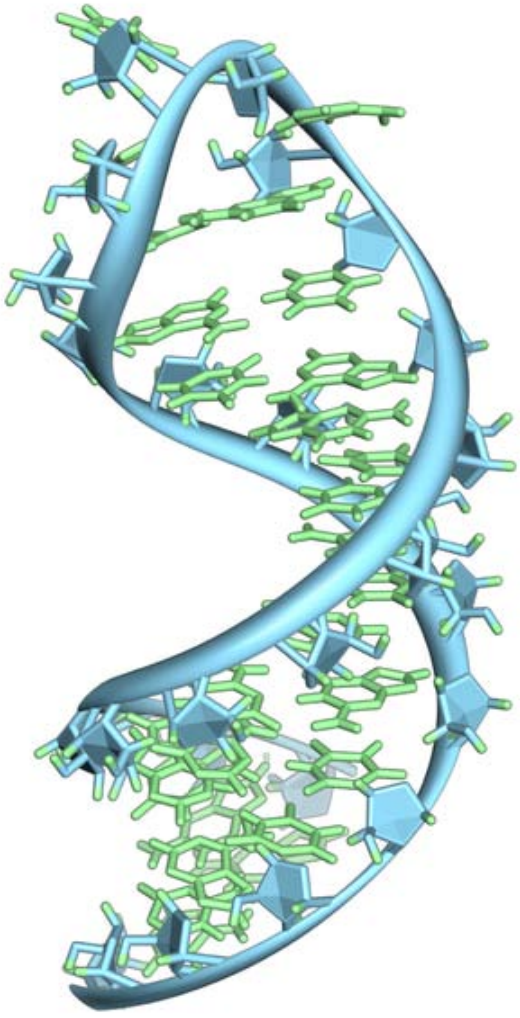
the information that the DNA bases are paired — also obtained through private communications from Erwin Chargaff in the previous years. Chargaff's rules played a very important role in establishing double-helix configurations for B-DNA as well as A-DNA.

Experimental evidence supporting the Watson and Crick model were published in a series of five articles in the same issue of *Nature*. Of these, Franklin and Gosling's paper was the first publication of their own X-ray diffraction data and original analysis method that partially supported the Watson and Crick model; this issue also contained an article on DNA structure by Maurice Wilkins and two of his colleagues, whose analysis and *in vivo* B-DNA X-ray patterns also supported the presence *in vivo* of the double-helical DNA configurations as proposed by Crick and Watson for their double-helix molecular model of DNA in the previous two pages of *Nature*. In 1962, after Franklin's death, Watson, Crick, and Wilkins jointly received the Nobel Prize in Physiology or Medicine. However, Nobel rules of the time allowed only living recipients, but a vigorous debate continues on who should receive credit for the discovery.

In an influential presentation in 1957, Crick laid out the central dogma of molecular biology, which foretold the relationship between DNA, RNA, and proteins, and articulated the "adaptor hypothesis". Final confirmation of the replication mechanism that was implied by the double-helical structure followed in 1958 through the Meselson–Stahl experiment. Further work by Crick and coworkers showed that the genetic code was based on non-overlapping triplets of bases, called codons, allowing Har Gobind Khorana, Robert W. Holley and Marshall Warren Nirenberg to decipher the genetic code. These findings represent the birth of molecular biology.

Chapter- 3

RNA



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

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

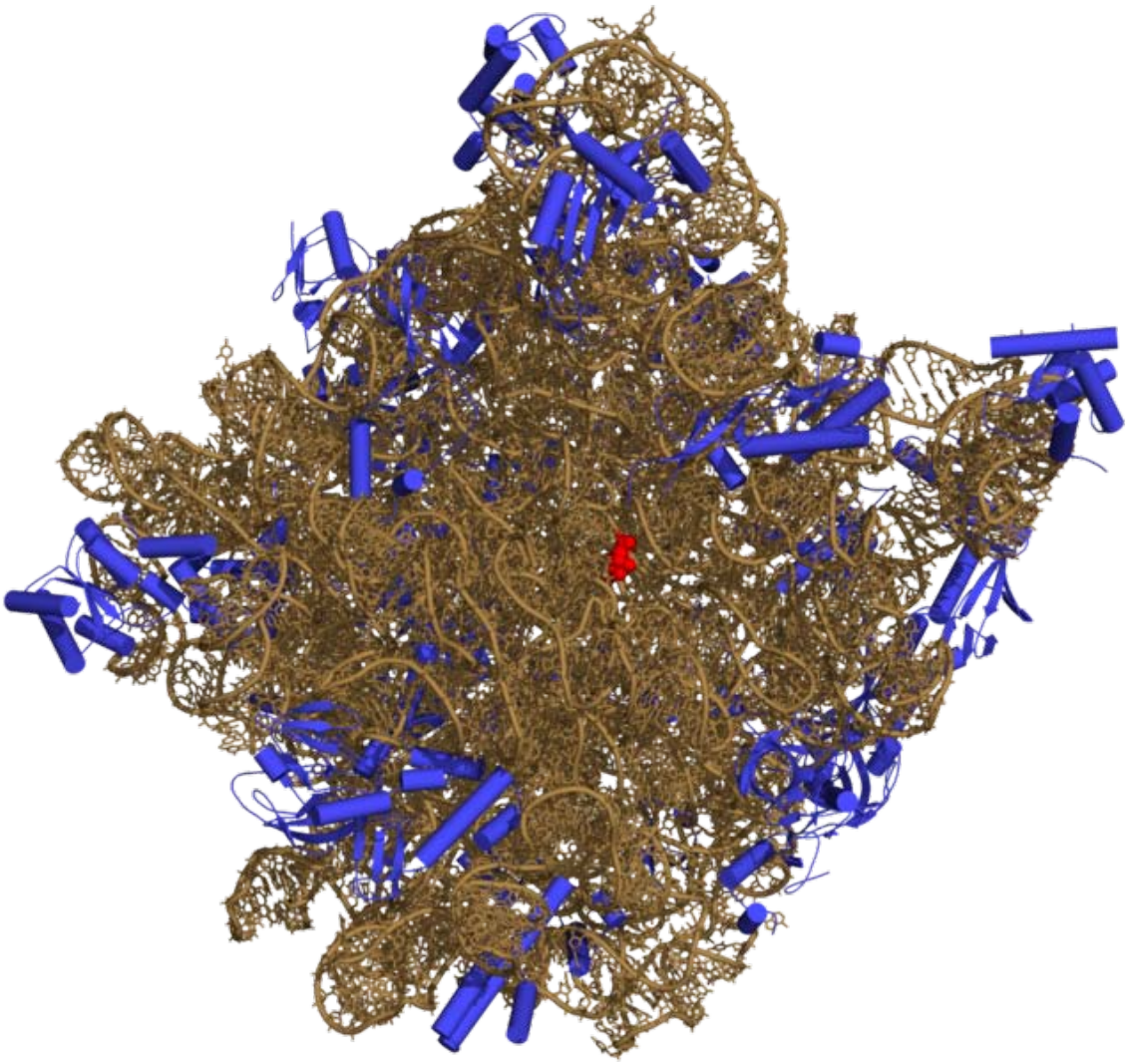
Like DNA, RNA is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase (sometimes called a nitrogenous base), a ribose sugar, and a phosphate group. The sequence of nucleotides allows RNA to encode genetic information. For example, some viruses use RNA instead of DNA as their genetic material, and all organisms use messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins.

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

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

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

Comparison with DNA



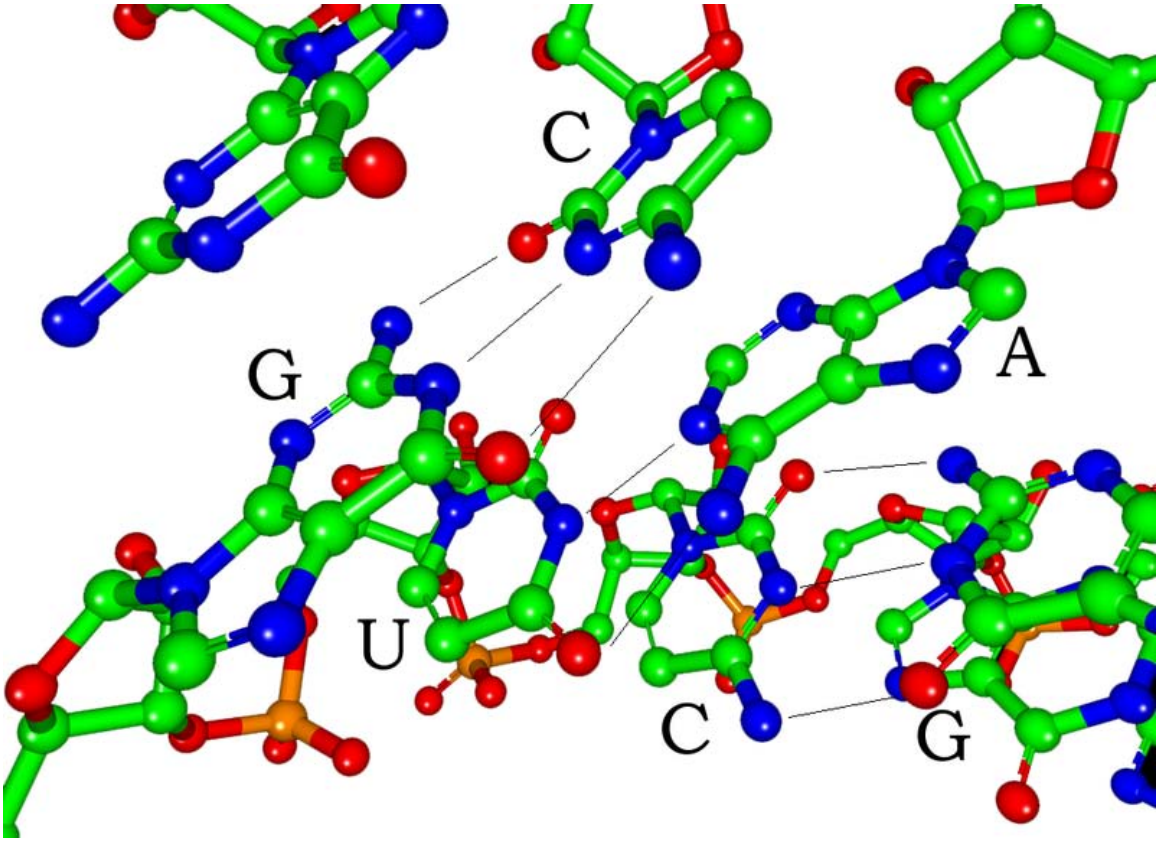
Three-dimensional representation of the 50S ribosomal subunit. RNA is in ochre, protein in blue. The active site is in the middle (red).

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

Like DNA, most biologically active RNAs, including mRNA, tRNA, rRNA, snRNAs, and other non-coding RNAs, contain self-complementary sequences that allow parts of the RNA to fold and pair with itself to form double helices. Structural analysis of these

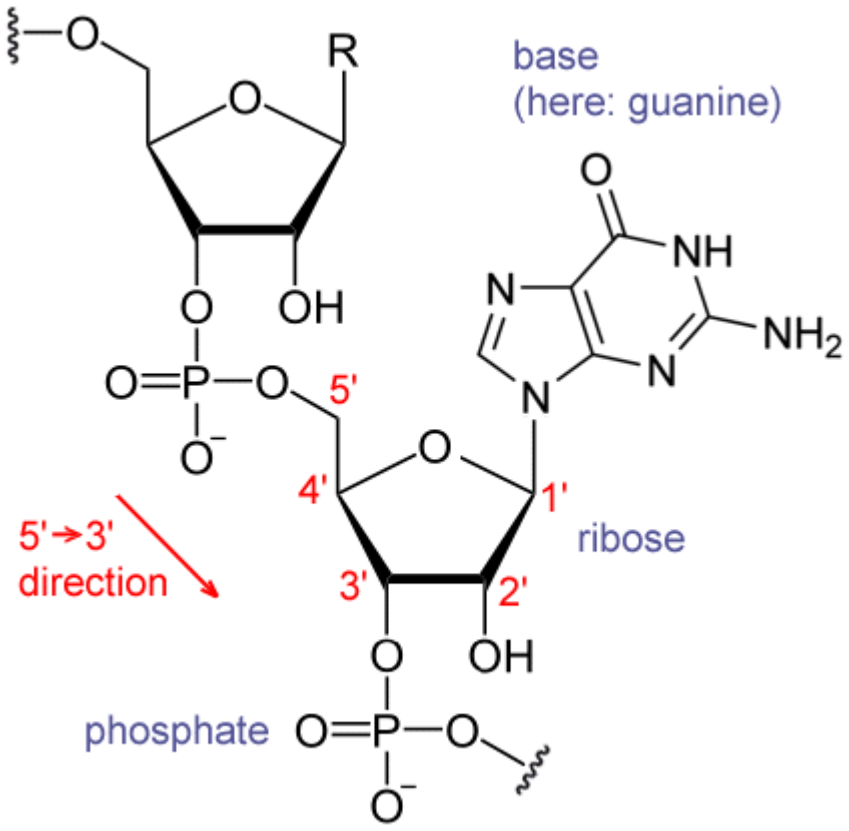
RNAs has revealed that they are highly structured. Unlike DNA, their structures do not consist of long double helices but rather collections of short helices packed together into structures akin to proteins. In this fashion, RNAs can achieve chemical catalysis, like enzymes. For instance, determination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

Structure



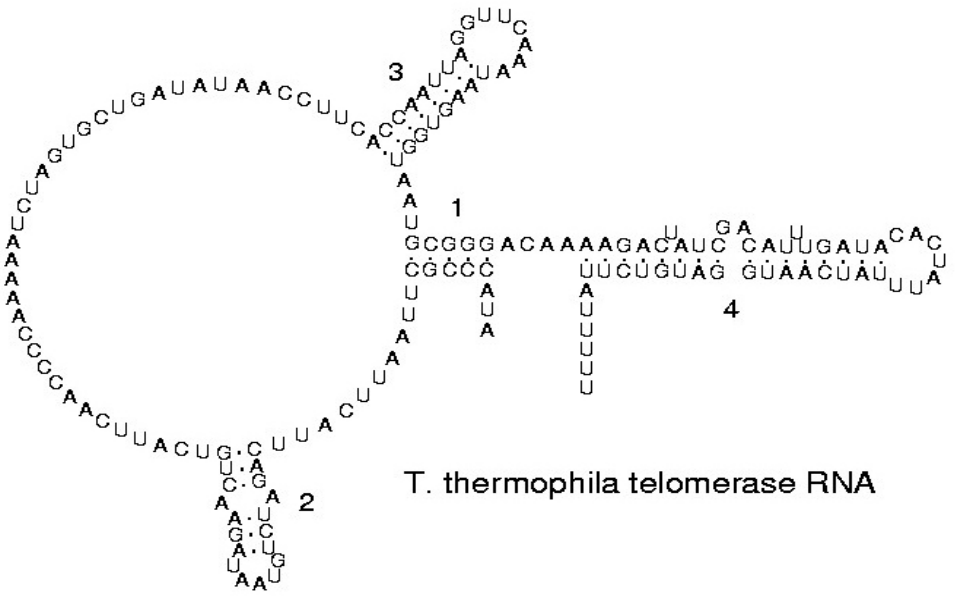
Watson-Crick base pairs in a siRNA (hydrogen atoms are not shown)

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



Chemical structure of RNA

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



Secondary structure of a telomerase RNA

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

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

The functional form of single stranded RNA molecules, just like proteins, frequently requires a specific tertiary structure. The scaffold for this structure is provided by secondary structural elements that are hydrogen bonds within the molecule. This leads to several recognizable "domains" of secondary structure like hairpin loops, bulges, and internal loops. Since RNA is charged, metal ions such as Mg^{2+} are needed to stabilise many secondary and tertiary structures.

Synthesis

Synthesis of RNA is usually catalyzed by an enzyme—RNA polymerase—using DNA as a template, a process known as transcription. Initiation of transcription begins with the

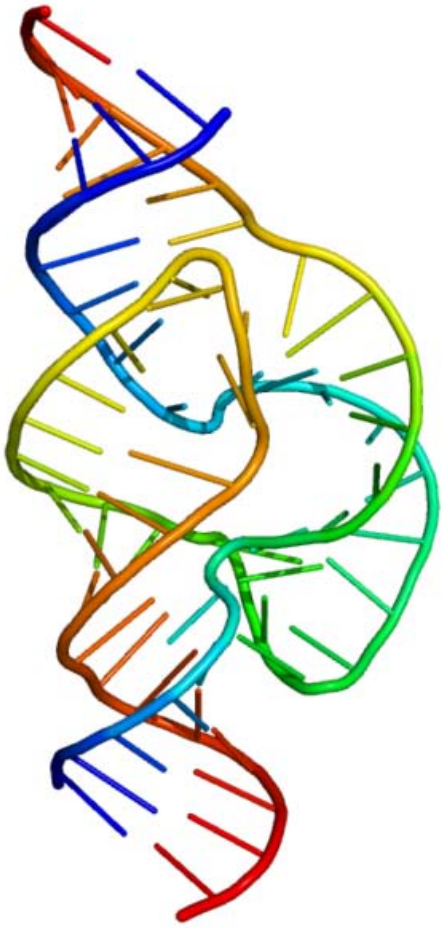
binding of the enzyme to a promoter sequence in the DNA (usually found "upstream" of a gene). The DNA double helix is unwound by the helicase activity of the enzyme. The enzyme then progresses along the template strand in the 3’ to 5’ direction, synthesizing a complementary RNA molecule with elongation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

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

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

Types of RNA

Overview



Structure of a hammerhead ribozyme, a ribozyme that cuts RNA

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

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

In translation

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

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

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

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

Regulatory RNAs

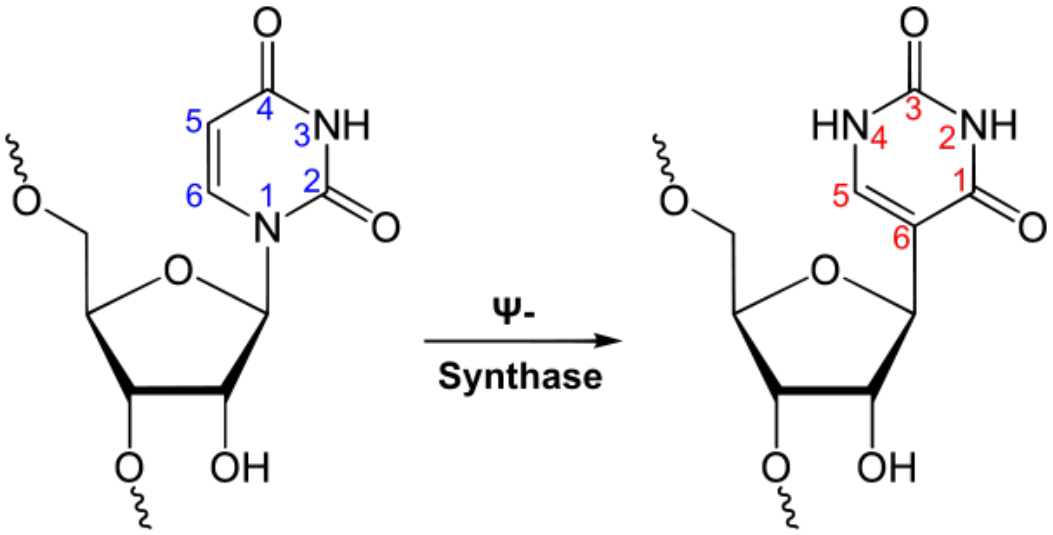
Several types of RNA can downregulate gene expression by being complementary to a part of an mRNA or a gene's DNA. MicroRNAs (miRNA; 21-22 nt) are found in eukaryotes and act through RNA interference (RNAi), where an effector complex of miRNA and enzymes can break down mRNA to which the miRNA is complementary, block the mRNA from being translated, or accelerate its degradation. While small interfering RNAs (siRNA; 20-25 nt) are often produced by breakdown of viral RNA, there are also endogenous sources of siRNAs.

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

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

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

In RNA processing



Uridine to pseudouridine is a common RNA modification

Many RNAs are involved in modifying other RNAs. Introns are spliced out of pre-mRNA by spliceosomes, which contain several small nuclear RNAs (snRNA), or the introns can be ribozymes that are spliced by themselves. RNA can also be altered by having its nucleotides modified to other nucleotides than A, C, G and U. In eukaryotes, modifications of RNA nucleotides are generally directed by small nucleolar RNAs (snoRNA; 60-300 nt), found in the nucleolus and cajal bodies. snoRNAs associate with enzymes and guide them to a spot on an RNA by basepairing to that RNA. These enzymes then perform the nucleotide modification. rRNAs and tRNAs are extensively modified, but snRNAs and mRNAs can also be the target of base modification.

RNA genomes

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

In reverse transcription

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

Double-stranded RNA

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

Key discoveries in RNA biology

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

The sequence of the 77 nucleotides of a yeast tRNA was found by Robert W. Holley in 1965, winning Holley the 1968 Nobel Prize in Medicine (shared with Har Gobind

Khorana and Marshall Nirenberg). In 1967, Carl Woese hypothesized that RNA might be catalytic and suggested that the earliest forms of life (self-replicating molecules) could have relied on RNA both to carry genetic information and to catalyze biochemical reactions—an RNA world.

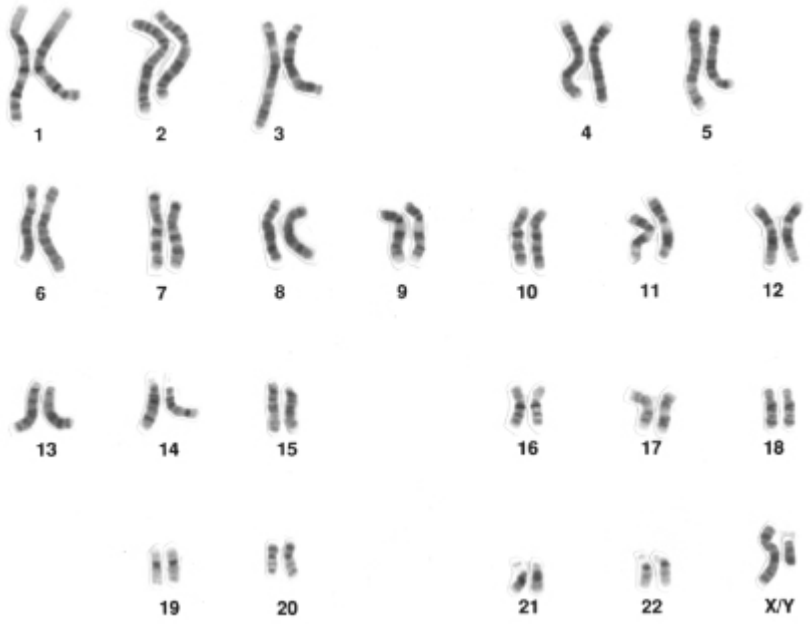
During the early 1970s, retroviruses and reverse transcriptase were discovered, showing for the first time that enzymes could copy RNA into DNA (the opposite of the usual route for transmission of genetic information). For this work, David Baltimore, Renato Dulbecco and Howard Temin were awarded a Nobel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

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

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

Chapter- 4

Genome



An image of the 46 chromosomes, making up the diploid genome of human male. (The mitochondrial chromosome is not shown.)

In modern molecular biology and genetics, the **genome** is the entirety of an organism's hereditary information. It is encoded either in DNA or, for many types of virus, in RNA. The genome includes both the genes and the non-coding sequences of the DNA.

Origin of Term

The term was adapted in 1920 by Hans Winkler, Professor of Botany at the University of Hamburg, Germany. In Greek, the word *genome* (γίνομαι) means "I become, I am born, to come into being". The Oxford English Dictionary suggests the name to be a blend of the words ***gene*** and *chromos**ome***. A few related *-ome* words already existed, such as *biome* and *rhizome*, forming a vocabulary into which *genome* fits systematically.

Overview

Some organisms have multiple copies of chromosomes, diploid, triploid, tetraploid and so on. In classical genetics, in a sexually reproducing organism (typically eukarya) the gamete has half of the number of chromosome of the somatic cell and the genome is a full set of chromosomes in a gamete. In haploid organisms, including cells of bacteria, archaea, and in organelles including mitochondria and chloroplasts, or viruses, that similarly contain genes, the single or set of circular and/or linear chains of DNA (or RNA for some viruses), likewise constitute the *genome*. The term genome can be applied specifically to mean that stored on a complete set of *nuclear DNA* (i.e., the "nuclear genome") but can also be applied to that stored within organelles that contain their own DNA, as with the "mitochondrial genome" or the "chloroplast genome". Additionally, the genome can comprise nonchromosomal genetic elements such as viruses, plasmids, and transposable elements.

When people say that the genome of a sexually reproducing species has been "sequenced", typically they are referring to a determination of the sequences of one set of autosomes and one of each type of sex chromosome, which together represent both of the possible sexes. Even in species that exist in only one sex, what is described as "a genome sequence" may be a composite read from the chromosomes of various individuals. In general use, the phrase "genetic makeup" is sometimes used conversationally to mean the genome of a particular individual or organism. The study of the global properties of genomes of related organisms is usually referred to as genomics, which distinguishes it from genetics which generally studies the properties of single genes or groups of genes.

Both the number of base pairs and the number of genes vary widely from one species to another, and there is only a rough correlation between the two (an observation known as the C-value paradox). At present, the highest known number of genes is around 60,000, for the protozoan causing trichomoniasis, almost three times as many as in the human genome.

An analogy to the human genome stored on DNA is that of instructions stored in a book:

- The book (genome) would contain 23 chapters (chromosomes);
- each chapter contains 48 to 250 million letters (A,C,G,T) without spaces;
- Hence, the book contains over 3.2 billion letters total;
- The book fits into a cell nucleus the size of a pinpoint;
- At least one copy of the book (all 23 chapters) is contained in every cell of our body.

Types

Most biological entities that are more complex than a virus sometimes or always carry additional genetic material besides that which resides in their chromosomes. In some contexts, such as sequencing the genome of a pathogenic microbe, "genome" is meant to include information stored on this auxiliary material, which is carried in plasmids. In

such circumstances then, "genome" describes all of the genes and information on non-coding DNA that have the potential to be present.

In eukaryotes such as plants, protozoa and animals, however, "genome" carries the typical connotation of only information on chromosomal DNA. So although these organisms contain chloroplasts and/or mitochondria that have their own DNA, the genetic information contained by DNA within these organelles is not considered part of the genome. In fact, mitochondria are sometimes said to have their own genome often referred to as the "mitochondrial genome". The DNA found within the chloroplast may be referred to as the "plastome".

Genomes and genetic variation

Note that a genome does not capture the genetic diversity or the genetic polymorphism of a species. For example, the human genome sequence in principle could be determined from just half the information on the DNA of one cell from one individual. To learn what variations in genetic information underlie particular traits or diseases requires comparisons across individuals. This point explains the common usage of "genome" (which parallels a common usage of "gene") to refer not to the information in any particular DNA sequence, but to a whole family of sequences that share a biological context.

Although this concept may seem counter intuitive, it is the same concept that says there is no particular shape that is the shape of a cheetah. Cheetahs vary, and so do the sequences of their genomes. Yet both the individual animals and their sequences share commonalities, so one can learn something about cheetahs and "cheetah-ness" from a single example of either.

Sequencing and mapping

The Human Genome Project was organized to map and to sequence the human genome. Other genome projects include mouse, rice, the plant *Arabidopsis thaliana*, the puffer fish, bacteria like E. coli, etc. In 1976, Walter Fiers at the University of Ghent (Belgium) was the first to establish the complete nucleotide sequence of a viral RNA-genome (bacteriophage MS2). The first DNA-genome project to be completed was the Phage Φ-X174, with only 5386 base pairs, which was sequenced by Fred Sanger in 1977. The first bacterial genome to be completed was that of Haemophilus influenzae, completed by a team at The Institute for Genomic Research in 1995.

The development of new technologies has dramatically decreased the difficulty and cost of sequencing, and the number of complete genome sequences is rising rapidly. Among many genome database sites, the one maintained by the US National Institutes of Health is inclusive.

These new technologies open up the prospect of personal genome sequencing as an important diagnostic tool. A major step toward that goal was the completion of the decipherment of the full genome of DNA pioneer James D. Watson in 2007.

Whereas a genome sequence lists the order of every DNA base in a genome, a genome map identifies the landmarks. A genome map is less detailed than a genome sequence and aids in navigating around the genome.

Comparison of different genome sizes

Organism type	Organism	Genome size (base pairs)	mass - in pg	Note
Virus	Bacteriophage MS2	3,569	0.000002	First sequenced RNA-genome
Virus	SV40	5,224		
Virus	Phage Φ-X174	5,386		First sequenced DNA-genome
Virus	HIV	9749		
Virus	Phage λ	48,502		
Virus	Mimivirus	1,181,404		Largest known viral genome
Bacterium	*Haemophilus influenzae*	1,830,000		First genome of a living organism sequenced, July 1995
Bacterium	*Carsonella ruddii*	159,662		Smallest non-viral genome.
Bacterium	*Buchnera aphidicola*	600,000		
Bacterium	*Wigglesworthia glossinidia*	700,000		
Bacterium	*Escherichia coli*	4,600,000		
Bacterium	*Solibacter usitatus* (strain Ellin 6076)	9,970,000		Largest known Bacterial genome
Amoeboid	*Polychaos dubium* (*"Amoeba" dubia*)	670,000,000,000	737	Largest known genome.
Plant	*Arabidopsis thaliana*	157,000,000		First plant genome sequenced, December 2000.
Plant	*Genlisea margaretae*	63,400,000		Smallest recorded flowering plant genome, 2006.
Plant	*Fritillaria assyrica*	130,000,000,000		
Plant	*Populus trichocarpa*	480,000,000		First tree genome sequenced, September

				2006
Plant	*Paris japonica* (Japanese-native, pale-petal)	150,000,000,000	152.23 pg	Largest plant genome known
Moss	*Physcomitrella patens*	480,000,000		First genome of a bryophyte sequenced, January 2008.
Yeast	*Saccharomyces cerevisiae*	12,100,000		
Fungus	*Aspergillus nidulans*	30,000,000		
Nematode	*Caenorhabditis elegans*	100,300,000		First multicellular animal genome sequenced, December 1998
Nematode	*Pratylenchus coffeae*	20,000,000		Smallest animal genome known
Insect	*Drosophila melanogaster* (fruit fly)	130,000,000		
Insect	*Bombyx mori* (silk moth)	530,000,000		
Insect	*Apis mellifera* (honey bee)	236,000,000		
Insect	*Solenopsis invicta* (fire ant)	480,000,000		
Fish	*Tetraodon nigroviridis* (type of puffer fish)	385,000,000		Smallest vertebrate genome known
Mammal	*Homo sapiens*	3,200,000,000	3	
Fish	*Protopterus aethiopicus* (marbled lungfish)	130,000,000,000	143	Largest vertebrate genome known

Note: The DNA from a single (diploid) human cell if the 46 chromosomes were connected end-to-end and straightened, would have a length of ~2 m and a width of ~2.4 nanometers.

Since genomes and their organisms are very complex, one research strategy is to reduce the number of genes in a genome to the bare minimum and still have the organism in question survive. There is experimental work being done on minimal genomes for single cell organisms as well as minimal genomes for multicellular organisms. The work is both *in vivo* and *in silico*.

Genome evolution

Genomes are more than the sum of an organism's genes and have traits that may be measured and studied without reference to the details of any particular genes and their products. Researchers compare traits such as *chromosome number* (karyotype), genome size, gene order, codon usage bias, and GC-content to determine what mechanisms could have produced the great variety of genomes that exist today.

Duplications play a major role in shaping the genome. Duplications may range from extension of short tandem repeats, to duplication of a cluster of genes, and all the way to duplications of entire chromosomes or even entire genomes. Such duplications are probably fundamental to the creation of genetic novelty.

Horizontal gene transfer is invoked to explain how there is often extreme similarity between small portions of the genomes of two organisms that are otherwise very distantly related. Horizontal gene transfer seems to be common among many microbes. Also, eukaryotic cells seem to have experienced a transfer of some genetic material from their chloroplast and mitochondrial genomes to their nuclear chromosomes.

Chapter- 5

Heredity

Heredity is the passing of traits to offspring (from its parent or ancestors). This is the process by which an offspring cell or organism acquires or becomes predisposed to the characteristics of its parent cell or organism. Through heredity, variations exhibited by individuals can accumulate and cause some species to evolve. The study of heredity in biology is called genetics, which includes the field of epigenetics.

History

The ancients had a variety of ideas about heredity: Theophrastus proposed that male flowers caused female flowers to ripen; Hippocrates speculated that "seeds" were produced by various body parts and transmitted to offspring at the time of conception, and Aristotle thought that male and female semen mixed at conception. Aeschylus, in 458 BC, proposed the male as the parent, with the female as a "nurse for the young life sown within her."

Various hereditary mechanisms were envisaged without being properly tested or quantified. These included blending inheritance and the inheritance of acquired traits. Nevertheless, people were able to develop domestic breeds of animals as well as crops through artificial selection. The inheritance of acquired traits also formed a part of early Lamarckian ideas on evolution.

In the 9th century AD, the Afro-Arab writer Al-Jahiz considered the effects of the environment on the likelihood of an animal to survive, and first described the struggle for existence. His ideas on the struggle for existence in the *Book of Animals* have been summarized as follows:

Animals engage in a struggle for existence; for resources, to avoid being eaten and to breed. Environmental factors influence organisms to develop new characteristics to ensure survival, thus transforming into new species. Animals that survive to breed can pass on their successful characteristics to offspring.

In 1000 AD, the Arab physician, Abu al-Qasim al-Zahrawi (known as Albucasis in the West), wrote the first clear description of haemophilia, a hereditary genetic disorder, in his *Al-Tasrif*. In this work, he wrote of an Andalusian family whose males died of bleeding after minor injuries.

During the 18th century, Dutch microscopist Antonie van Leeuwenhoek (1632–1723) discovered "animalcules" in the sperm of humans and other animals. Some scientists speculated they saw a "little man" (homunculus) inside each sperm. These scientists formed a school of thought known as the "spermists." They contended the only contributions of the female to the next generation were the womb in which the homunculus grew, and prenatal influences of the womb. An opposing school of thought, the ovists, believed that the future human was in the egg, and that sperm merely stimulated the growth of the egg. Ovists thought women carried eggs containing boy and girl children, and that the gender of the offspring was determined well before conception.

Types of heredity

Dominant and recessive

An allele is said to be dominant if it is always expressed in the appearance of an organism (phenotype). For example, in peas the allele for green pods, G, is dominant to that for yellow pods, g. Since the allele for green pods is dominant, pea plants with the pair of alleles GG (homozygote) or Gg (heterozygote) will have green pods. The allele for yellow pods is recessive. The effects of this allele are only seen when it is present in both chromosomes, gg (homozygote).

The description of a mode of biological inheritance consists of three main categories:

1. Number of involved loci

- Monogenetic (also called "simple") – one locus
- Oligogenetic – few loci
- Polygenetic – many loci

2. Involved chromosomes

- Autosomal – loci are not situated on a sex chromosome
- Gonosomal – loci are situated on a sex chromosome
    - X-chromosomal – loci are situated on the X chromosome (the more common case)
    - Y-chromosomal – loci are situated on the Y chromosome
- Mitochondrial – loci are situated on the mitochondrial DNA

3. Correlation genotype–phenotype

- Dominant

- Intermediate (also called "codominant")
- Recessive

These three categories are part of every exact description of a mode of inheritance in the above order. Additionally, more specifications may be added as follows:

4. Coincidental and environmental interactions

- Penetrance
  - Complete
  - Incomplete (percentual number)
- Expressivity
  - Invariable
  - Variable
- Heritability (in polygenetic and sometimes also in oligogenetic modes of inheritance)
- Maternal or paternal imprinting phenomena

5. Sex-linked interactions

- Sex-linked inheritance (gonosomal loci)
- Sex-limited phenotype expression (e.g., cryptorchism)
- Inheritance through the maternal line (in case of mitochondrial DNA loci)
- Inheritance through the paternal line (in case of Y-chromosomal loci)

6. Locus–locus interactions

- Epistasis with other loci (e.g., overdominance)
- Gene coupling with other loci
- Homozygotous lethal factors
- Semi-lethal factors

Determination and description of a mode of inheritance is primarily achieved through statistical analysis of pedigree data. In case the involved loci are known, methods of molecular genetics can also be employed.

Charles Darwin: theory of evolution

When Charles Darwin proposed his theory of evolution in 1859, one of its major problems was the lack of an underlying mechanism for heredity. Darwin believed in a mix of blending inheritance and the inheritance of acquired traits (pangenesis). Blending inheritance would lead to uniformity across populations in only a few generations and thus would remove variation from a population on which natural selection could act. This led to Darwin adopting some Lamarckian ideas in later editions of *On the Origin of Species* and his later biological works. Darwin's primary approach to heredity was to outline how it appeared to work (noticing that traits could be inherited which were not

expressed explicitly in the parent at the time of reproduction, that certain traits could be sex-linked, etc.) rather than suggesting mechanisms.

Darwin's initial model of heredity was adopted by, and then heavily modified by, his cousin Francis Galton, who laid the framework for the biometric school of heredity. Galton rejected the aspects of Darwin's pangenesis model which relied on acquired traits.

The inheritance of acquired traits was shown to have little basis in the 1880s when August Weismann cut the tails off many generations of mice and found that their offspring continued to develop tails.

Gregor Mendel: father of modern genetics

The idea of particulate inheritance of genes can be attributed to the Moravian monk Gregor Mendel who published his work on pea plants in 1865. However, his work was not widely known and was rediscovered in 1901. It was initially assumed the Mendelian inheritance only accounted for large (qualitative) differences, such as those seen by Mendel in his pea plants—and the idea of additive effect of (quantitative) genes was not realised until R.A. Fisher's (1918) paper, "The Correlation Between Relatives on the Supposition of Mendelian Inheritance."

Modern development of genetics and heredity

In the 1930s, work by Fisher and others resulted in a combination of Mendelian and biometric schools into the modern evolutionary synthesis. The modern synthesis bridged the gap between experimental geneticists and naturalists; and between both and palaeontologists, stating that:

1. All evolutionary phenomena can be explained in a way consistent with known genetic mechanisms and the observational evidence of naturalists.
2. Evolution is gradual: small genetic changes, recombination ordered by natural selection. Discontinuities amongst species (or other taxa) are explained as originating gradually through geographical separation and extinction (not saltation).
3. Selection is overwhelmingly the main mechanism of change; even slight advantages are important when continued. The object of selection is the phenotype in its surrounding environment. The role of genetic drift is equivocal; though strongly supported initially by Dobzhansky, it was downgraded later as results from ecological genetics were obtained.
4. The primacy of population thinking: the genetic diversity carried in natural populations is a key factor in evolution. The strength of natural selection in the wild was greater than expected; the effect of ecological factors such as niche occupation and the significance of barriers to gene flow are all important.
5. In palaeontology, the ability to explain historical observations by extrapolation from micro to macro-evolution is proposed. Historical contingency means

explanations at different levels may exist. Gradualism does not mean constant rate of change.

The idea that speciation occurs after populations are reproductively isolated has been much debated. In plants, polyploidy must be included in any view of speciation. Formulations such as 'evolution consists primarily of changes in the frequencies of alleles between one generation and another' were proposed rather later. The traditional view is that developmental biology ('evo-devo') played little part in the synthesis, but an account of Gavin de Beer's work by Stephen Jay Gould suggests he may be an exception.

Almost all aspects of the synthesis have been challenged at times, with varying degrees of success. There is no doubt, however, that the synthesis was a great landmark in evolutionary biology. It cleared up many confusions, and was directly responsible for stimulating a great deal of research in the post-World War II era.

Trofim Lysenko however caused a backlash of what is now called Lysenkoism in the Soviet Union when he emphasised Lamarckian ideas on the inheritance of acquired traits. This movement affected agricultural research and led to food shortages in the 1960s and seriously affected the USSR.

Chapter- 6

Mutation

In molecular biology and genetics, **mutations** are changes in a genomic sequence: the DNA sequence of a cell's genome or the DNA or RNA sequence of a virus. Mutations are caused by radiation, viruses, transposons and mutagenic chemicals, as well as errors that occur during meiosis or DNA replication. They can also be induced by the organism itself, by cellular processes such as hypermutation.

Mutation can result in several different types of change in DNA sequences; these can either have no effect, alter the product of a gene, or prevent the gene from functioning properly or completely. Studies in the fly *Drosophila melanogaster* suggest that if a mutation changes a protein produced by a gene, this will probably be harmful, with about 70 percent of these mutations having damaging effects, and the remainder being either neutral or weakly beneficial. Due to the damaging effects that mutations can have on genes, organisms have mechanisms such as DNA repair to remove mutations.

Therefore, the optimal mutation rate for a species is a trade-off between costs of a high mutation rate, such as deleterious mutations, and the metabolic costs of maintaining systems to reduce the mutation rate, such as DNA repair enzymes. Viruses that use RNA as their genetic material have rapid mutation rates, which can be an advantage since these viruses will evolve constantly and rapidly, and thus evade the defensive responses of e.g. the human immune system.

Description

Mutations can involve large sections of DNA becoming duplicated, usually through genetic recombination. These duplications are a major source of raw material for evolving new genes, with tens to hundreds of genes duplicated in animal genomes every million years. Most genes belong to larger families of genes of shared ancestry. Novel genes are produced by several methods, commonly through the duplication and mutation of an ancestral gene, or by recombining parts of different genes to form new combinations with new functions.

Here, domains act as modules, each with a particular and independent function, that can be mixed together to produce genes encoding new proteins with novel properties. For example, the human eye uses four genes to make structures that sense light: three for color vision and one for night vision; all four arose from a single ancestral gene. Another advantage of duplicating a gene (or even an entire genome) is that this increases redundancy; this allows one gene in the pair to acquire a new function while the other copy performs the original function. Other types of mutation occasionally create new genes from previously noncoding DNA.

Changes in chromosome number may involve even larger mutations, where segments of the DNA within chromosomes break and then rearrange. For example, two chromosomes in the *Homo* genus fused to produce human chromosome 2; this fusion did not occur in the lineage of the other apes, and they retain these separate chromosomes. In evolution, the most important role of such chromosomal rearrangements may be to accelerate the divergence of a population into new species by making populations less likely to interbreed, and thereby preserving genetic differences between these populations.

Sequences of DNA that can move about the genome, such as transposons, make up a major fraction of the genetic material of plants and animals, and may have been important in the evolution of genomes. For example, more than a million copies of the Alu sequence are present in the human genome, and these sequences have now been recruited to perform functions such as regulating gene expression. Another effect of these mobile DNA sequences is that when they move within a genome, they can mutate or delete existing genes and thereby produce genetic diversity.

A mutation has caused this garden moss rose to produce flowers of different colors. This is a somatic mutation that may also be passed on in the germ line.

In multicellular organisms with dedicated reproductive cells, mutations can be subdivided into germ line mutations, which can be passed on to descendants through their reproductive cells, and somatic mutations (also called acquired mutations), which involve cells outside the dedicated reproductive group and which are not usually transmitted to descendants. If the organism can reproduce asexually through mechanisms such as cuttings or budding the distinction can become blurred.

For example, plants can sometimes transmit somatic mutations to their descendants asexually or sexually where flower buds develop in somatically mutated parts of plants. A new mutation that was not inherited from either parent is called a *de novo* mutation.

The source of the mutation is unrelated to the consequence, although the consequences are related to which cells were mutated.

Nonlethal mutations accumulate within the gene pool and increase the amount of genetic variation. The abundance of some genetic changes within the gene pool can be reduced by natural selection, while other "more favorable" mutations may accumulate and result in adaptive changes.

For example, a butterfly may produce offspring with new mutations. The majority of these mutations will have no effect; but one might change the color of one of the butterfly's offspring, making it harder (or easier) for predators to see. If this color change is advantageous, the chance of this butterfly surviving and producing its own offspring are a little better, and over time the number of butterflies with this mutation may form a larger percentage of the population.

Neutral mutations are defined as mutations whose effects do not influence the fitness of an individual. These can accumulate over time due to genetic drift. It is believed that the overwhelming majority of mutations have no significant effect on an organism's fitness. Also, DNA repair mechanisms are able to mend most changes before they become permanent mutations, and many organisms have mechanisms for eliminating otherwise permanently mutated somatic cells.

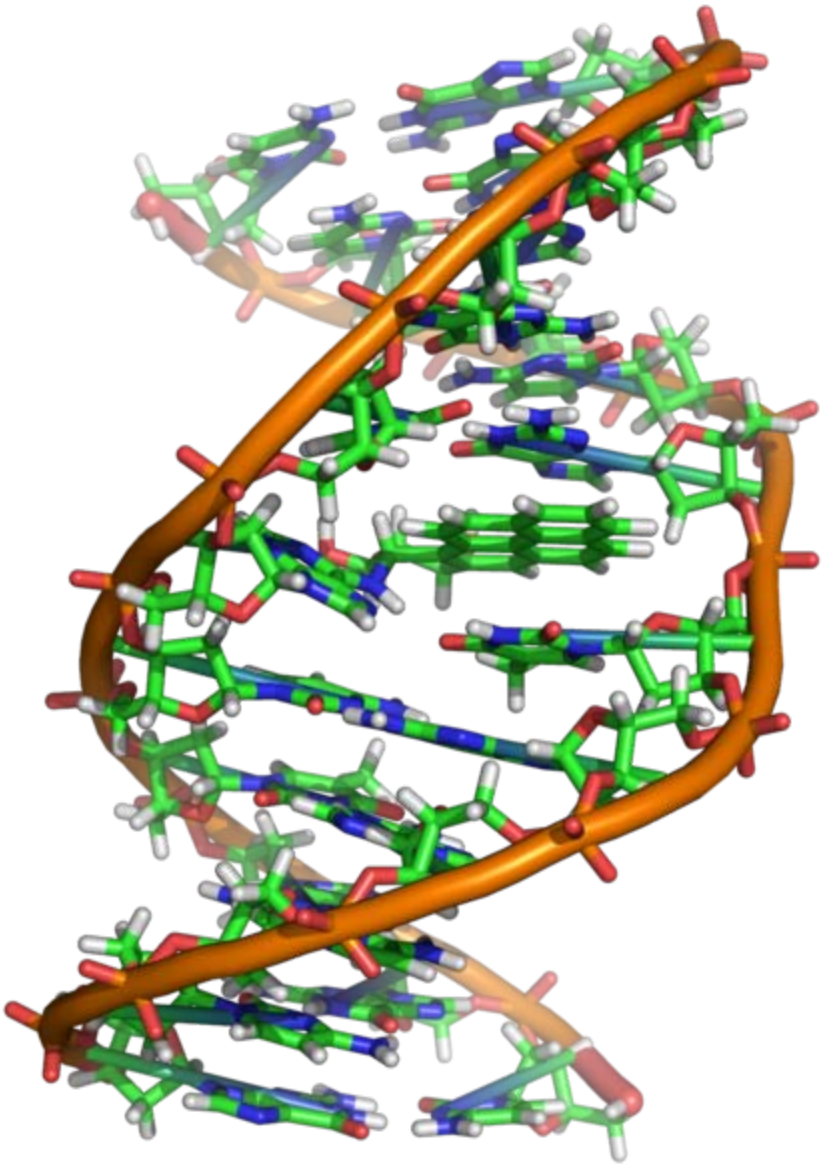
Mutation is generally accepted by biologists as the mechanism by which natural selection acts, generating advantageous new traits that survive and multiply in offspring as well as disadvantageous traits, in less fit offspring, that tend to die out.

Causes

Two classes of mutations are spontaneous mutations (molecular decay) and induced mutations caused by mutagens.

Spontaneous mutations on the molecular level can be caused by:

- Tautomerism – A base is changed by the repositioning of a hydrogen atom, altering the hydrogen bonding pattern of that base resulting in incorrect base pairing during replication.
- Depurination – Loss of a purine base (A or G) to form an apurinic site (AP site).
- Deamination – Hydrolysis changes a normal base to an atypical base containing a keto group in place of the original amine group. Examples include C → U and A → HX (hypoxanthine), which can be corrected by DNA repair mechanisms; and 5MeC (5-methylcytosine) → T, which is less likely to be detected as a mutation because thymine is a normal DNA base.
- Slipped strand mispairing - Denaturation of the new strand from the template during replication, followed by renaturation in a different spot ("slipping"). This can lead to insertions or deletions.



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

Induced mutations on the molecular level can be caused by:

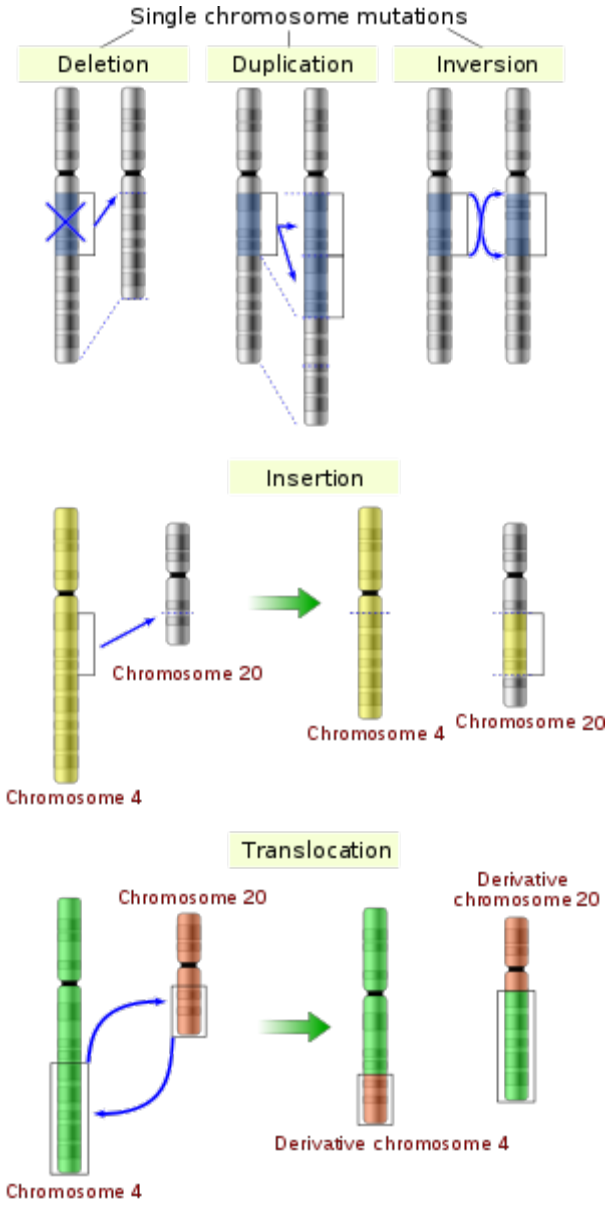
- Chemicals
  - Hydroxylamine NH_2OH
  - Base analogs (e.g. BrdU)
  - Alkylating agents (e.g. *N*-ethyl-*N*-nitrosourea) These agents can mutate both replicating and non-replicating DNA. In contrast, a base analog can only mutate the DNA when the analog is incorporated in replicating the DNA. Each of these classes of chemical mutagens has certain effects that then lead to transitions, transversions, or deletions.

    - Agents that form DNA adducts (e.g. ochratoxin A metabolites)
    - DNA intercalating agents (e.g. ethidium bromide)
    - DNA crosslinkers
    - Oxidative damage
    - Nitrous acid converts amine groups on A and C to diazo groups, altering their hydrogen bonding patterns which leads to incorrect base pairing during replication.
- Radiation
    - Ultraviolet radiation (nonionizing radiation). Two nucleotide bases in DNA – cytosine and thymine – are most vulnerable to radiation that can change their properties. UV light can induce adjacent pyrimidine bases in a DNA strand to become covalently joined as a pyrimidine dimer. UV radiation, particularly longer-wave UVA, can also cause oxidative damage to DNA.
    - Ionizing radiation
    - Radioactive decay, such as ^{14}C in DNA
- Viral infections

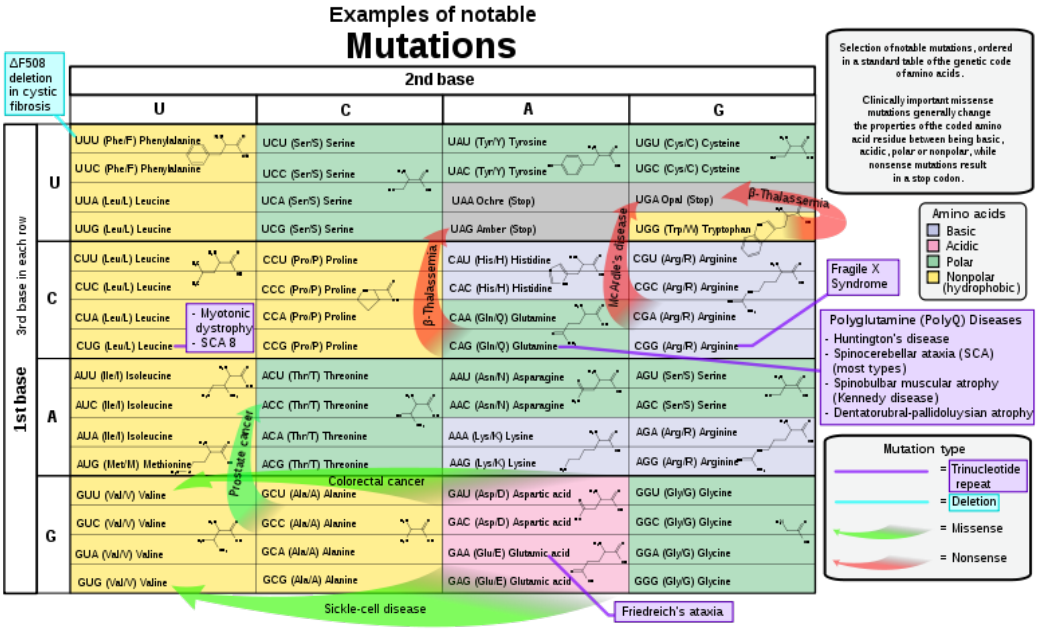
DNA has so-called hotspots, where mutations occur up to 100 times more frequently than the normal mutation rate. A hotspot can be at an unusual base, e.g., 5-methylcytosine.

Mutation rates also vary across species. Evolutionary biologists have theorized that higher mutation rates are beneficial in some situations, because they allow organisms to evolve and therefore adapt more quickly to their environments. For example, repeated exposure of bacteria to antibiotics, and selection of resistant mutants, can result in the selection of bacteria that have a much higher mutation rate than the original population (mutator strains).

Classification of mutation types



Illustrations of five types of chromosomal mutations



Selection of disease-causing mutations, in a standard table of the genetic code of amino acids

By effect on structure

The sequence of a gene can be altered in a number of ways. Gene mutations have varying effects on health depending on where they occur and whether they alter the function of essential proteins. Mutations in the structure of genes can be classified as:

- Small-scale mutations, such as those affecting a small gene in one or a few nucleotides, including:
  - **Point mutations**, often caused by chemicals or malfunction of DNA replication, exchange a single nucleotide for another. These changes are classified as transitions or transversions. Most common is the transition that exchanges a purine for a purine (A ↔ G) or a pyrimidine for a pyrimidine, (C ↔ T). A transition can be caused by nitrous acid, base mis-pairing, or mutagenic base analogs such as 5-bromo-2-deoxyuridine (BrdU). Less common is a transversion, which exchanges a purine for a pyrimidine or a pyrimidine for a purine (C/T ↔ A/G). An example of a transversion is adenine (A) being converted into a cytosine (C). A point mutation can be reversed by another point mutation, in which the nucleotide is changed back to its original state (true reversion) or by second-site reversion (a complementary mutation elsewhere that results in regained gene functionality). Point mutations that occur within the protein coding region of a gene may be classified into three kinds, depending upon what the erroneous codon codes for:
    - Silent mutations: which code for the same amino acid.
    - Missense mutations: which code for a different amino acid.

        - Nonsense mutations: which code for a stop and can truncate the protein.
    - **Insertions** add one or more extra nucleotides into the DNA. They are usually caused by transposable elements, or errors during replication of repeating elements (e.g. AT repeats). Insertions in the coding region of a gene may alter splicing of the mRNA (splice site mutation), or cause a shift in the reading frame (frameshift), both of which can significantly alter the gene product. Insertions can be reverted by excision of the transposable element.
    - **Deletions** remove one or more nucleotides from the DNA. Like insertions, these mutations can alter the reading frame of the gene. They are generally irreversible: though exactly the same sequence might theoretically be restored by an insertion, transposable elements able to revert a very short deletion (say 1–2 bases) in *any* location are either highly unlikely to exist or do not exist at all. Note that a deletion is not the exact opposite of an insertion: the former is quite random while the latter consists of a specific sequence inserting at locations that are not entirely random or even quite narrowly defined.
- Large-scale mutations in chromosomal structure, including:
    - **Amplifications** (or gene duplications) leading to multiple copies of all chromosomal regions, increasing the dosage of the genes located within them.
    - **Deletions** of large chromosomal regions, leading to loss of the genes within those regions.
    - Mutations whose effect is to juxtapose previously separate pieces of DNA, potentially bringing together separate genes to form functionally distinct fusion genes (e.g. bcr-abl). These include:
        - **Chromosomal translocations**: interchange of genetic parts from nonhomologous chromosomes.
        - **Interstitial deletions**: an intra-chromosomal deletion that removes a segment of DNA from a single chromosome, thereby apposing previously distant genes. For example, cells isolated from a human astrocytoma, a type of brain tumor, were found to have a chromosomal deletion removing sequences between the "fused in glioblastoma" (fig) gene and the receptor tyrosine kinase "ros", producing a fusion protein (FIG-ROS). The abnormal FIG-ROS fusion protein has constitutively active kinase activity that causes oncogenic transformation (a transformation from normal cells to cancer cells).
        - **Chromosomal inversions**: reversing the orientation of a chromosomal segment.
    - **Loss of heterozygosity**: loss of one allele, either by a deletion or recombination event, in an organism that previously had two different alleles.

By effect on function

- **Loss-of-function mutations** are the result of gene product having less or no function. When the allele has a complete loss of function (null allele) it is often called an **amorphic mutation**. Phenotypes associated with such mutations are most often recessive. Exceptions are when the organism is haploid, or when the reduced dosage of a normal gene product is not enough for a normal phenotype (this is called haploinsufficiency).
- **Gain-of-function mutations** change the gene product such that it gains a new and abnormal function. These mutations usually have dominant phenotypes. Often called a neomorphic mutation.
- **Dominant negative mutations** (also called **antimorphic mutations**) have an altered gene product that acts antagonistically to the wild-type allele. These mutations usually result in an altered molecular function (often inactive) and are characterised by a dominant or semi-dominant phenotype. In humans, Marfan syndrome is an example of a dominant negative mutation occurring in an autosomal dominant disease. In this condition, the defective glycoprotein product of the fibrillin gene (FBN1) antagonizes the product of the normal allele.
- **Lethal mutations** are mutations that lead to the death of the organisms which carry the mutations.
- A **back mutation** or **reversion** is a point mutation that restores the original sequence and hence the original phenotype.

By effect on fitness

In applied genetics it is usual to speak of mutations as either harmful or beneficial.

- A **harmful mutation** is a mutation that decreases the fitness of the organism.
- A **beneficial mutation** is a mutation that increases fitness of the organism, or which promotes traits that are desirable.

In theoretical population genetics, it is more usual to speak of such mutations as deleterious or advantageous. In the neutral theory of molecular evolution, genetic drift is the basis for most variation at the molecular level.

- A **neutral mutation** has no harmful or beneficial effect on the organism. Such mutations occur at a steady rate, forming the basis for the molecular clock.
- A **deleterious mutation** has a negative effect on the phenotype, and thus decreases the fitness of the organism.
- An **advantageous mutation** has a positive effect on the phenotype, and thus increases the fitness of the organism.
- A **nearly neutral mutation** is a mutation that may be slightly deleterious or advantageous, although most nearly neutral mutations are slightly deleterious.

In reality, viewing the fitness effects of mutations in these discrete categories is an oversimplification. Attempts have been made to infer the distribution of fitness effects

using mutagenesis experiments or theoretical models applied to molecular sequence data. However, the current distribution is still uncertain, and some aspects of the distribution likely vary between species.

By inheritance

- inheritable generic in pro-generic tissue or cells on path to be changed to gametes.
- non inheritable **somatic** (e.g., carcinogenic mutation)
- non inheritable post mortem aDNA mutation in decaying remains.

By pattern of inheritance The human genome contains two copies of each gene – a paternal and a maternal allele.

- A **heterozygous mutation** is a mutation of only one allele.
- A **homozygous mutation** is an identical mutation of both the paternal and maternal alleles.
- **Compound heterozygous** mutations or a **genetic compound** comprises two different mutations in the paternal and maternal alleles.
- A **wildtype** or **homozygous non-mutated** organism is one in which neither allele is mutated. (Just not a mutation)

By impact on protein sequence

- A **frameshift mutation** is a mutation caused by insertion or deletion of a number of nucleotides that is not evenly divisible by three from a DNA sequence. Due to the triplet nature of gene expression by codons, the insertion or deletion can disrupt the reading frame, or the grouping of the codons, resulting in a completely different translation from the original. The earlier in the sequence the deletion or insertion occurs, the more altered the protein produced is.

- A **nonsense mutation** is a point mutation in a sequence of DNA that results in a premature stop codon, or a *nonsense codon* in the transcribed mRNA, and possibly a truncated, and often nonfunctional protein product.

- **Missense mutations** or *nonsynonymous mutations* are types of point mutations where a single nucleotide is changed to cause substitution of a different amino acid. This in turn can render the resulting protein nonfunctional. Such mutations are responsible for diseases such as Epidermolysis bullosa, sickle-cell disease, and SOD1 mediated ALS (Boillée 2006, p. 39).

- A **neutral mutation** is a mutation that occurs in an amino acid codon which results in the use of a different, but chemically similar, amino acid. The similarity between the two is enough that little or no change is often rendered in the protein. For example, a change from AAA to AGA will encode arginine, a chemically similar molecule to the intended lysine. Neutral mutations occur because of the degenerate nature of the genetic code.

- **Silent mutations** are mutations that do not result in a change to the amino acid sequence of a protein. They may occur in a region that does not code for a protein, or they may occur within a codon in a manner that does not alter the final amino acid sequence. The phrase *silent mutation* is often used interchangeably with the phrase *synonymous mutation*; however, synonymous mutations are a subcategory of the former, occurring only within exons. The name silent could be a misnomer. For example, a silent mutation in the exon/intron border may lead to alternative splicing by changing the splice site, thereby leading to a changed protein.

Special classes

- **Conditional mutation** is a mutation that has wild-type (or less severe) phenotype under certain "permissive" environmental conditions and a mutant phenotype under certain "restrictive" conditions. For example, a temperature-sensitive mutation can cause cell death at high temperature (restrictive condition), but might have no deleterious consequences at a lower temperature (permissive condition).

Nomenclature

A committee of the Human Genome Variation Society (HGVS) has developed the standard human sequence variant nomenclature, which should be used by researchers and DNA diagnostic centers to generate unambiguous mutation descriptions. In principle, this nomenclature can also be used to describe mutations in other organisms. The nomenclature specifies the type of mutation and base or amino acid changes.

- Nucleotide substitution (e.g. 76A>T) - The number is the position of the nucleotide from the 5' end, the first letter represents the wild type nucleotide, and the second letter represents the nucleotide which replaced the wild type. In the given example, the adenine at the 76th position was replaced by a thymine.
    - If it becomes necessary to differentiate between mutations in genomic DNA, mitochondrial DNA, and RNA, a simple convention is used. For example, if the 100th base of a nucleotide sequence mutated from G to C, then it would be written as g.100G>C if the mutation occurred in genomic DNA, m.100G>C if the mutation occurred in mitochondrial DNA, or r.100g>c if the mutation occurred in RNA. Note that for mutations in RNA, the nucleotide code is written in lower case.
- Amino acid substitution (e.g. D111E) – The first letter is the one letter code of the wild type amino acid, the number is the position of the amino acid from the N terminus, and the second letter is the one letter code of the amino acid present in the mutation. Nonsense mutations are represented with an X for the second amino acid (e.g. D111X).
- Amino acid deletion (e.g. ΔF508) – The Greek letter Δ (delta) indicates a deletion. The letter refers to the amino acid present in the wild type and the number is the position from the N terminus of the amino acid were it to be present as in the wild type.

The complete set of rules and more examples of mutation descriptions can be found at the HGVS sequence variant nomenclature website. Since the nomenclature has to cover all sequence variants, descriptions can become very complex. To prevent mistakes and facilitate correct use of this nomenclature, the journal Human Mutation recommends the use of Mutalyzer, which can apply the HGVS human nomenclature guidelines to check and, if necessary, correct sequence variant descriptions.

Harmful mutations

Changes in DNA caused by mutation can cause errors in protein sequence, creating partially or completely non-functional proteins. To function correctly, each cell depends on thousands of proteins to function in the right places at the right times. When a mutation alters a protein that plays a critical role in the body, a medical condition can result. A condition caused by mutations in one or more genes is called a genetic disorder. Some mutations alter a gene's DNA base sequence but do not change the function of the protein made by the gene. Studies of the fly *Drosophila melanogaster* suggest that if a mutation does change a protein, this will probably be harmful, with about 70 percent of these mutations having damaging effects, and the remainder being either neutral or weakly beneficial. However, studies in yeast have shown that only 7% of mutations that are not in genes are harmful.

If a mutation is present in a germ cell, it can give rise to offspring that carries the mutation in all of its cells. This is the case in hereditary diseases. On the other hand, a mutation may occur in a somatic cell of an organism. Such mutations will be present in all descendants of this cell within the same organism, and certain mutations can cause the cell to become malignant, and thus cause cancer.

Often, gene mutations that could cause a genetic disorder are repaired by the DNA repair system of the cell. Each cell has a number of pathways through which enzymes recognize and repair mistakes in DNA. Because DNA can be damaged or mutated in many ways, the process of DNA repair is an important way in which the body protects itself from disease.

Beneficial mutations

Although most mutations that change protein sequences are neutral or harmful, some mutations have a positive effect on an organism. In this case, the mutation may enable the mutant organism to withstand particular environmental stresses better than wild-type organisms, or reproduce more quickly. In these cases a mutation will tend to become more common in a population through natural selection.

For example, a specific 32 base pair deletion in human CCR5 (CCR5-Δ32) confers HIV resistance to homozygotes and delays AIDS onset in heterozygotes. The CCR5 mutation is more common in those of European descent. One possible explanation of the etiology of the relatively high frequency of CCR5-Δ32 in the European population is that it conferred resistance to the bubonic plague in mid-14th century Europe. People with this

mutation were more likely to survive infection; thus its frequency in the population increased. This theory could explain why this mutation is not found in southern Africa, where the bubonic plague never reached. A newer theory suggests that the selective pressure on the CCR5 Delta 32 mutation was caused by smallpox instead of the bubonic plague.

Another example, is Sickle cell disease which is a blood disorder in which the body produces an abnormal type of the oxygen-carrying substance hemoglobin in the red blood cells. One-third of all indigenous inhabitants of Sub-Saharan Africa carry the gene, because in areas where malaria is common, there is a survival value in carrying only a single sickle-cell gene (sickle cell trait). Those with only one of the two alleles of the sickle-cell disease are more resistant to malaria, since the infestation of the malaria plasmodium is halted by the sickling of the cells which it infests.

Prion mutation

Prions are proteins and do not contain genetic material. However, prion replication has been shown to be subject to mutation and natural selection just like other forms of replication.

Chapter- 7

Nucleotide and Genetic Variation

Nucleotide

Nucleotides are molecules that, when joined together, make up the structural units of RNA and DNA. In addition, nucleotides play central roles in metabolism. In that capacity, they serve as sources of chemical energy (adenosine triphosphate and guanosine triphosphate), participate in cellular signaling (cyclic guanosine monophosphate and cyclic adenosine monophosphate), and are incorporated into important cofactors of enzymatic reactions (coenzyme A, flavin adenine dinucleotide, flavin mononucleotide, and nicotinamide adenine dinucleotide phosphate).

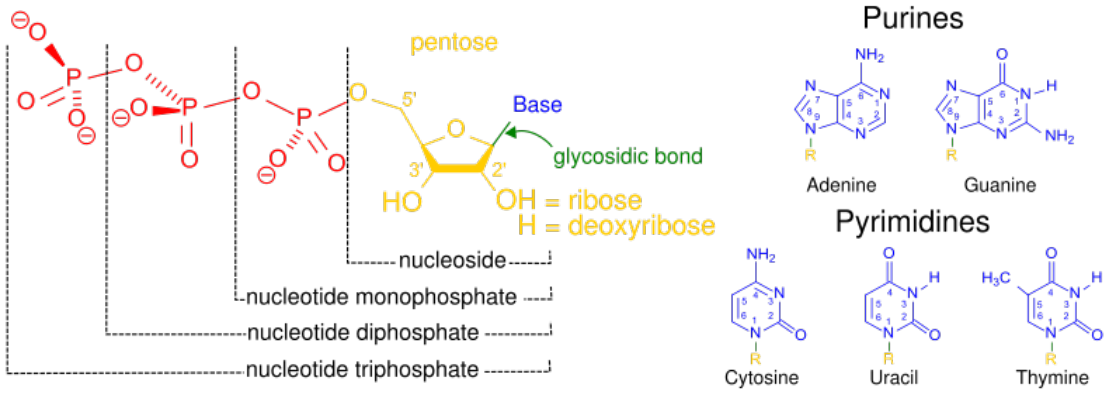


Figure 1: Structural elements of the most common nucleotides

Nucleotide structure

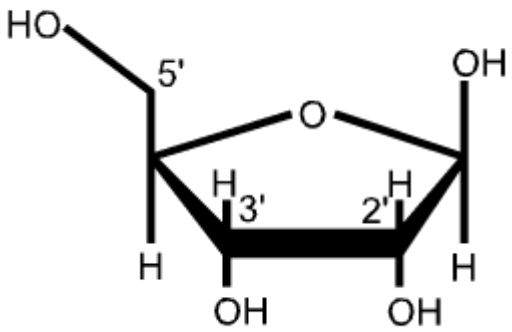


Figure 2: Ribose structure indicating numbering of carbon atoms

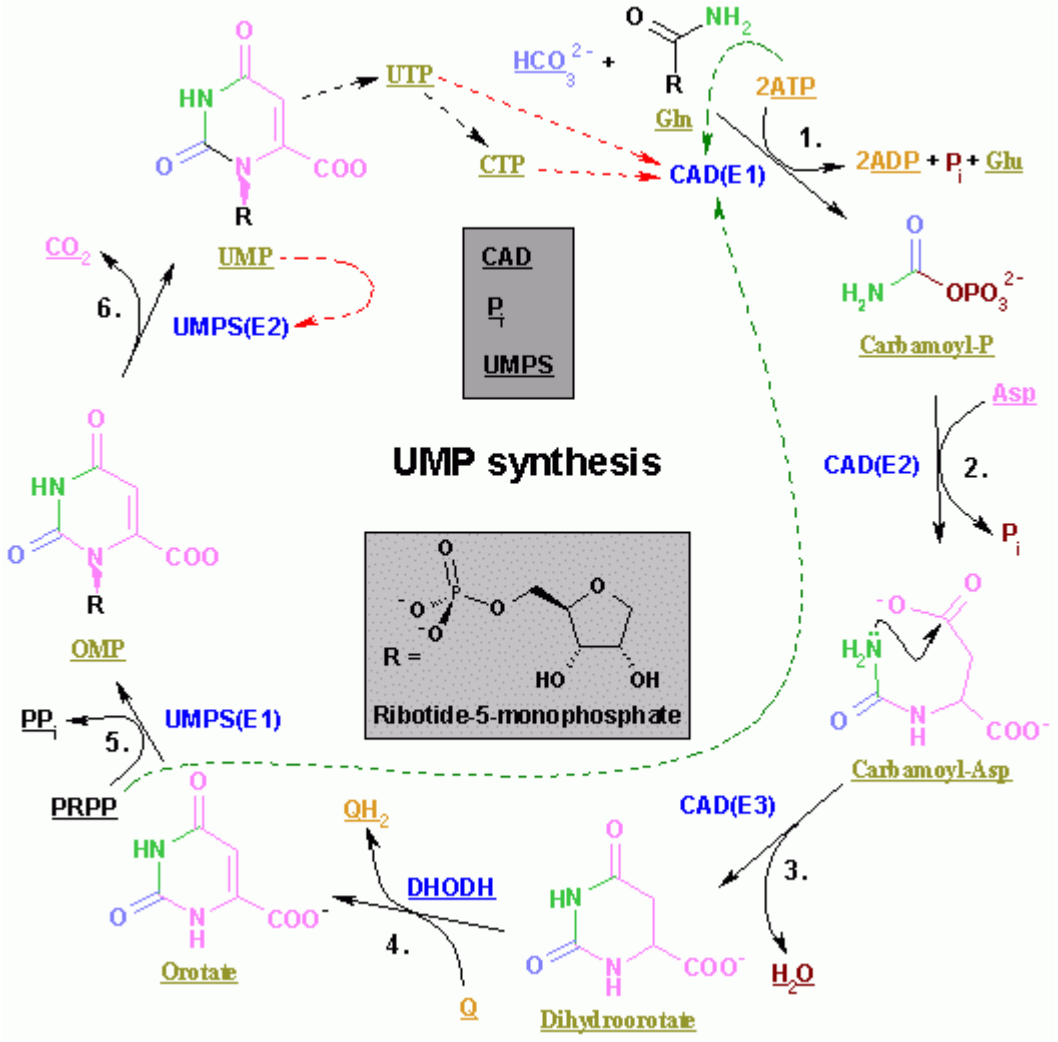
A nucleotide is composed of a nucleobase (nitrogenous base), a five-carbon sugar (either ribose or 2'-deoxyribose), and one to three phosphate groups. Together, the nucleobase and sugar comprise a nucleoside. The phosphate groups form bonds with either the 2, 3, or 5-carbon of the sugar, with the 5-carbon site most common. Cyclic nucleotides form when the phosphate group is bound to two of the sugar's hydroxyl groups. Ribonucleotides are nucleotides where the sugar is ribose, and deoxyribonucleotides contain the sugar deoxyribose. Nucleotides can contain either a purine or a pyrimidine base.

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, each due to their unique structures.

Synthesis

Nucleotides can be synthesized by a variety of means both in vitro and in vivo. In vivo, nucleotides can be synthesised de novo or recycled through salvage pathways. Nucleotides undergo breakdown such that useful parts can be reused in synthesis reactions to create new nucleotides. In vitro, protecting groups may be used during laboratory production of nucleotides. A purified nucleoside is protected to create a phosphoramidite, which can then be used to obtain analogues not found in nature and/or to synthesize an oligonucleotide.

Pyrimidine ribonucleotides

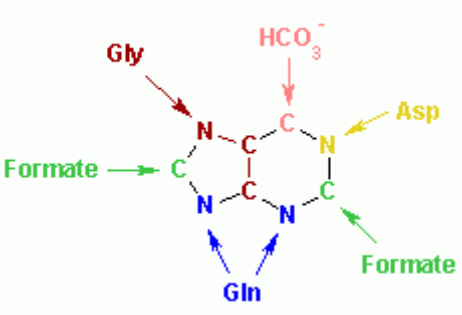


The synthesis of UMP.
The color scheme is as follows: **enzymes, coenzymes, substrate names, inorganic molecules**

Pyrimidine nucleotide synthesis starts with the formation of carbamoyl phosphate from glutamine and CO_2. The cyclisation reaction between carbamoyl phosphate reacts with aspartate, yielding orotate in subsequent steps. Orotate reacts with 5-phosphoribosyl α-diphosphate (PRPP), yielding orotidine monophosphate (OMP), which is decarboxylated to form uridine monophosphate (UMP). It is from UMP that other pyrimidine nucleotides are derived. UMP is phosphorylated to uridine triphosphate (UTP) via two sequential reactions with ATP. Cytidine monophosphate (CMP) is derived from conversion of UTP to cytidine triphosphate (CTP) with subsequent loss of two phosphates.

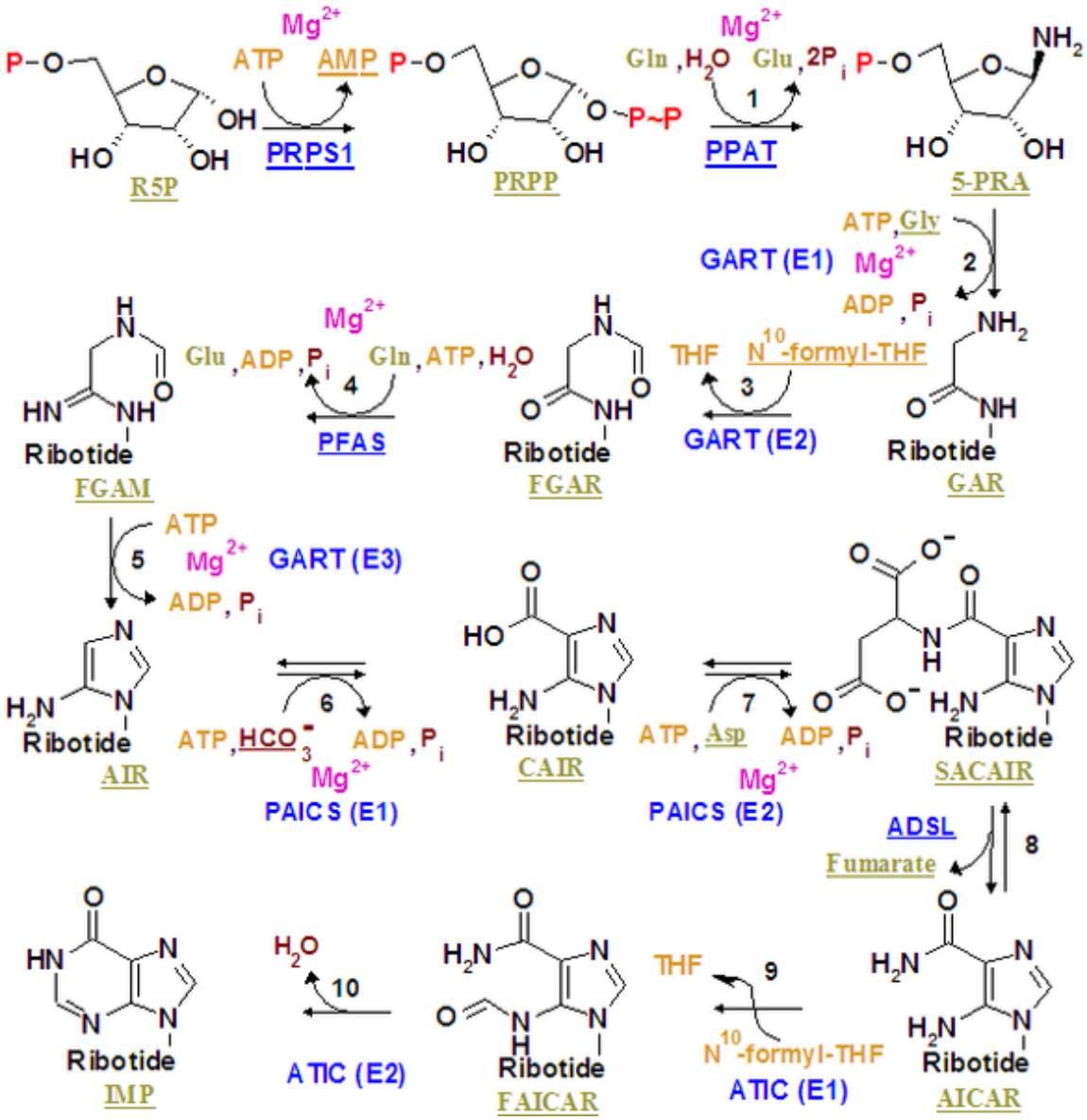
Purine ribonucleotides

The atoms which are used to build the purine nucleotides come from a variety of sources:



The biosynthetic origins of purine ring atoms

N1 arises from the amine group of Asp
C2 and C8 originate from formate
N3 and N9 are contributed by the amide group of Gln
C4, C5 and N7 are derived from Gly
C6 comes from HCO_3^- (CO_2)



The synthesis of IMP. The color scheme is as follows: **enzymes, coenzymes, substrate names, metal ions, inorganic molecules**

The de novo synthesis of purine nucleotides by which these precursors are incorporated into the purine ring proceeds by a 10-step pathway to the branch-point intermediate IMP, the nucleotide of the base hypoxanthine. AMP and GMP are subsequently synthesized from this intermediate via separate, two-step pathways. Thus, purine moieties are initially formed as part of the ribonucleotides rather than as free bases.

Six enzymes take part in IMP synthesis. Three of them are multifunctional:

- GART (reactions 2, 3, and 5)
- PAICS (reactions 6, and 7)
- ATIC (reactions 9, and 10)

Reaction 1. The pathway starts with the formation of PRPP. PRPS1 is the enzyme that activates R5P, which is formed primarily by the pentose phosphate pathway, to PRPP by reacting it with ATP. The reaction is unusual in that a pyrophosphoryl group is directly transferred from ATP to C1 of R5P and that the product has the **α** configuration about C1. This reaction is also shared with the pathways for the synthesis of the pyrimidine nucleotides, Trp, and His. As a result of being on (a) such (a) major metabolic crossroad and the use of energy, this reaction is highly regulated.

Reaction 2. In the first reaction unique to purine nucleotide biosynthesis, PPAT catalyzes the displacement of PRPP's pyrophosphate group (PP_i) by Gln's amide nitrogen. The reaction occurs with the inversion of configuration about ribose C1, thereby forming **β**-5-phosphorybosylamine (5-PRA) and establishing the anomeric form of the future nucleotide. This reaction, which is driven to completion by the subsequent hydrolysis of the released PP_i, is the pathway's flux-generating step and is therefore regulated, too.

Length unit

Nucleotide (abbreviated nt) is a common length unit for single-stranded RNA, similar to how base pair is a length unit for double-stranded DNA.

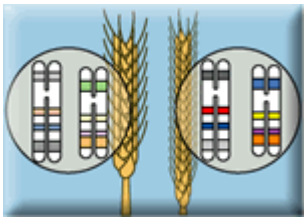
Abbreviation codes for degenerate bases

The IUPAC has designated the symbols for nucleotides. Apart from the five (A, G, C, T/U) bases, often degenerate bases are used especially for designing PCR primers. These nucleotide codes are listed here.

IUPAC nucleotide code	**Base**
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G

Y	C or T (U)
S	G or C
W	A or T (U)
K	G or T (U)
M	A or C
B	C or G or T (U)
D	A or G or T (U)
H	A or C or T (U)
V	A or C or G
N	any base
. or -	gap

Genetic variation



Genetic variation, variation in alleles of genes, occurs both within and among populations. Genetic variation is important because it provides the "raw material" for natural selection. Genetic variation are brought about by mutation,a change in a chemical structure of a gene. Polyploidy is a example of chromosomal mutation. Polyploidy is a condition wherein organisms has three or more set of genetic variation(3n or more).

Among individuals within a population

Genetic variation among individuals within a population can be identified at a variety of levels. It is possible to identify genetic variation from observations of phenotypic variation in either quantitative traits (traits that vary continuously and are coded for by many genes, e.g., leg length in dogs) or discrete traits (traits that fall into discrete categories and are coded for by one or a few genes (e.g., white, pink, red petal color in certain flowers).

Genetic variation can also be identified by examining variation at the level of enzymes using the process of protein electrophoresis. Polymorphic genes have more than one

allele at each locus. Half of the genes that code for enzymes in insects and plants may be polymorphic, whereas polymorphisms are less common in vertebrates.

Ultimately, genetic variation is caused by variation in the order of bases in the nucleotides in genes. New technology now allows scientists to directly sequence DNA which has identified even more genetic variation than was previously detected by protein electrophoresis. Examination of DNA has shown genetic variation in both coding regions and in the non-coding intron region of genes.

Genetic variation will result in phenotypic variation if variation in the order of nucleotides in the DNA sequence results in a difference in the order of amino acids in proteins coded by that DNA sequence, and if the resultant differences in amino acid sequence influence the shape, and thus the function of the enzyme.

Between populations

Geographic variation in genes often occurs among populations living in different locations. Geographic variation may be due to differences in selective pressures or to genetic drift.

Measurement

Genetic variation within a population is commonly measured as the percentage of gene loci that are polymorphic or the percentage of gene loci in individuals that are heterozygous.

Sources

Mutations are the ultimate source of genetic variation because they alter the order of bases in the nucleotides of DNA. Mutations are likely to be rare and most mutations are neutral or deleterious, but in some instances the new alleles can be favored by natural selection.

Genetic variation can also be produced by the recombination of chromosomes that occurs during sexual reproduction, called independent assortment.

Crossing over and random segregation during meiosis can result in the production of new alleles or new combinations of alleles. Furthermore, random ferilisation also contributes to variation.

Variation and recombination can be fascilitated by transposable and transposed genetic elements, commonly known as endogenous retroviruses, LINEs, SINEs, etc (see: variation-inducing genetic elements.)

Maintenance in populations

A variety of factors maintain genetic variation in populations. Potentially harmful recessive alleles can be hidden from selection in the heterozygous individuals in populations of diploid organisms (recessive alleles are only expressed in the less common homozygous individuals). Natural selection can also maintain genetic variation in balanced polymorphisms. Balanced polymorphisms may occur when heterozygotes are favored or when selection is frequency dependent.