

An Introduction to Molecular Biology

An Introduction to Molecular Biology

PHILIP MCFADDEN

OREGON STATE UNIVERSITY
CORVALLIS, OR



An Introduction to Molecular Biology by Philip McFadden is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License, except where otherwise noted.

Download for free at [https://open.oregonstate.education/
molecular-biology/](https://open.oregonstate.education/molecular-biology/)

Publication and on-going maintenance of this textbook is possible due to grant support from Oregon State University Ecampus.

This book was produced with Pressbooks (<https://pressbooks.com>) and rendered with Prince.

Contents

The Scope of Molecular Biology	1
1. Cells	5
2. Genomes	11
3. DNA and RNA molecules	31
4. The central dogma: Replication and expression of genomes	49
5. DNA Replication and Repair	57
6. 5. Molecular biology information tools	84
7. 4. Social aspects of molecular biology	88
8. Credits	89
9. Macromolecules and Cells	90
10. *Nucleus	147
11. *Cell Cycle	191
12. *DNA the unit of life	216
13. *Replication of DNA and its repair	270
14. *RNA:The ribonucleic acid	318
15. *Transcription of RNA and its modification	360
16. *Genetic Code	403
17. *Gene Expression	418
18. *Protein synthesis	462
19. *Function and structure of Proteins	498
*Quiz time	561
*Glossary	571

The Scope of Molecular Biology

This is not just science. It's also about making choices.

Molecular biologists are curious about how biological traits from past generations are inherited and expressed in the present generation and then passed on to the next generation. In this regard, molecular biology is really no different from “traditional” biology. The same questions are asked: How do we develop from a newly fertilized egg? How do we grow? How do we crawl, walk, slither, swim, or fly just the way our ancestors did? How is it that littermates differ, and how did one end up with the mother’s eyes and one with the father’s tail? Molecular biology seeks answers to these traditional questions, but at the molecular level. Our text will give an overview of cells and genomes, including how much DNA it takes to serve as the blueprint for an organism.

When we think in molecular terms, we need to take into account those tiniest of explanatory features, the molecules we know from chemistry. So we will do just that. Molecules are described in terms of their atoms, most commonly including carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and phosphorous (P). Groups of atoms that can be attached in varying locations are given names such as methyl (CH_3), amino (NH_2), carboxyl (COOH), and hydroxyl (OH). Whole molecules are given technically formal names such as 6-aminopurine or 5-methyluracil, which, like many named molecules, have become better known by their respective common names, adenine and thymine. Abbreviated names of molecules, such as A for adenine and T for thymine, make it easier to demonstrate our ideas of how molecules assemble into more

complicated arrangements. dATP, for example, refers to deoxyadenosine triphosphate, a building block of DNA chains that puts an A (adenine) into place. dTTP, or deoxythymidine triphosphate, is the building block that puts a T (thymine) into place. As you may know, two complementary DNA chains associate with each other through interaction between their A's and T's (and their G's and C's). If you are not already familiar with chemistry, the conventions, abbreviations and common naming systems will generally allow you to gloss over many chemical details while zeroing in on fascinating explanations of how biology works at the molecular level.

At the most basic level, a genome can be written as a series of the four letters A, G, C and T. Deciphering the meaning of all those letters is the challenge. To proceed beyond DNA sequences we must first come to accept the notion that with few exceptions, each molecule in a living organism has a purpose. We can safely assume there is a reason why it is there. Molecular biologists aim to discover those reasons, to account for the inheritance and expression of biological traits. Fortunately, what is learned from one organism often applies equally well to another. Our text presents these common truths in the concise statement known as The Central Dogma, the series of information transforming steps that operate in living cells of every variety. All cells copy (*replicate*) their DNA sequences in about the same way. All cells write (*transcribe*) RNA sequences from DNA sequences in about the same way. All cells decode (*translate*) protein sequences from RNA sequences in about the same way. The codes involved are universal, so once the codes were cracked the science of molecular biology was assured of a long future in deciphering the treasure troves of information contained in every living organism.

While this intellectual pursuit has been largely successful and continues to make discoveries at a rapid pace, one of the most surprising outcomes has been the invention of new molecular tools as a side effect of the discoveries. Our text therefore will focus on this inventive side of molecular biology by describing some of the

most used and most important tools for analyzing molecular biology information. Here's a prime example. The fundamental biological question of how DNA is inherited led to the discovery of DNA copying enzymes known as DNA polymerases. These enzymes are now being used as highly versatile tools for copying and sequencing DNA for many purposes in technology and society. The polymerase chain reaction (PCR) that you have likely heard about for its ability to amplify small quantities of DNA is one such tool. In turn, PCR has made it possible to ask new biological questions such as how similar were Neanderthals to modern humans. The same technology has also been put to use to ask more immediate questions, such as medical questions of personal disease susceptibilities, legal questions of inheritance or paternity, and forensic questions such as whether an individual may have been present at a crime scene. This is just one illustration. Time and again, the discoveries of molecular biology have yielded new tools that have brought forth new and unexpected powers. Studying molecular biology can therefore teach us how life naturally works as well as how new life technologies are coming forward today and tomorrow.

With new this new molecular power to change cells and organisms, society faces many questions and choices. Science has definitely done this before. New and powerful human capabilities brought to the forefront by observations refined through testing and experimentation have often changed the world and our human history. When fire was first used to forge metals, perhaps the initiating idea was to fashion a better spearhead. Where will we now head with the new tools of molecular biology? This open question invites predictions and demands decisions when choices need to be made. Our text will present a number of open questions for us to ponder. Science can help address these questions, and an engineering perspective as to which designs are possible can also help provide guidance. But science and engineering are not sufficient. Coming up with answers is necessarily a synthesis of much more, including what we want and what we do not want. Informed decision makers will likely disagree on which answers are

the “right” answers. We can be certain that the future will bring some interesting choices. The debates on these matters are current and fervent. One way to keep track of the arguments is to assemble statements of the impact that an advance in molecular biology may have on people, society, the earth, and its many nonhuman denizens. We might not be able to resolve controversies, but laying out the opinions and making estimates of impact may be a prudent and positive approach. May we find the best way forward.

References and online resources

As a teacher and scientist at Oregon State University in the Department of Biochemistry and Biophysics, I have had the pleasure of having Kevin Ahern and Indira Rajagopal as my just-down-the-hallway neighbors since the 1990’s. I have learned a lot from them, and I highly recommend and applaud their pioneering efforts to make free textbooks available for students. I have taught general biochemistry from their *Biochemistry Free for All* over many academic terms. Therefore, for this current *Introduction to Molecular Biology* that I am giving to my BB 331 students, I am choosing to list their work first and foremost as outstanding reference material. In particular, in coming chapters, I will lift some of their illustrations and text from Chapters 7 and 8. So here you are, for free and in their entirety: Chapter 7 – Information Processing, and Chapter 8 – Basic Techniques, from Kevin and Indira, in easily readable formatting provided by LibreText.

I. Cells

The living cell is the unit of life

Let's go forward with the idea that by studying molecules we can learn more about biology. Now, which system shall we study at the molecular level? Are we curious about whales or whirligig beetles? Or if we are interested in lifestyle and energetics, maybe a comparison study – of sloths and hummingbirds – is in order. Perhaps we ought to examine the expression of traits in just one part of an organism, such as the stomata of the saguaro cactus that only opens at night to let in the carbon dioxide that the plant cell will turn into sugar (its too hot in Arizona to open the stomata during the day).

One way of making questions experimentally addressable is to narrow down the system of study into simpler units. As questions become more focused, we became more likely to discover the molecules that are part of the answers. Probably the greatest step forward in this regard was the recognition that the cell is the simplest unit that can generally be considered to be a living thing. If we ask questions at the cellular level, answers promptly begin to make sense at the molecular level, too. So let's review the kinds of cells that are distributed across the many branches of the tree of life.

Some organisms are unicellular, dividing into two separate cells at times in order to increase the population size. Unicellular organisms, which are generally invisible to the naked eye due to their small size (in the range of $1\mu\text{m}$ to $50\mu\text{m}$ in diameter, where $1\mu\text{m}$ or 1 micron is equal to one-millionth of a meter). Unicellular organisms include the true bacteria (classified as *prokaryotes*), the archaebacteria (also called archeons), and a wide variety of free-

living single cells such as yeast and protozoans that contain a cell nucleus (and are therefore classified as *eukaryotes*).

Though single cell organisms commonly survive independently of their brethren, such cells may also group together to form colonies, some of which involve elaborate communications between cells and separation of duties to help the colony survive. This theme of a separation of duties cells is fundamental to multicellular organisms, and has been taken to exquisite extremes in the organization of organisms such as ourselves. In multicellular organisms, specialized cells are arranged in groups to form elaborate tissues and organs responsible for major aspects of survival. Green plants have both roots and leaves. Sea urchins have both spines and ovaries. To its prey, T. rex may have looked like it was all teeth, but it also had a heart.

A typical cell diameter in a multicellular organism is 10 μm . A typical cell mass is 1 nanogram (one billionth of a gram), but these numbers vary widely. Take our nervous system as an example. Whereas the neurons and the glial cells in the gray matter of the brain frequently land among the average to smaller sized cells of our body, some neurons of the spinal cord have diameters in excess of 100 μm and may have fine extensions called axons that are long enough to extend from the lower brain stem to the fingers and toes.

In all, a human is made up of about 100 trillion or 10^{14} cells. One of the principal discoveries of molecular biology has been to verify that despite the differences between these many cells, they all contain essentially the full instruction set of genetic information that is inherited when sperm meets egg upon conception of a particular human being. During the development of the embryo, the fetus, the child, and the adult, many changes occur to differentiate one kind of cell from another. But the full set of genetic information remains present in each cell. That instruction set is the *genome*.

In eukaryotes, the genome is contained in the cell

nucleus

The nucleus is a membrane enclosed organelle found in eukaryotic cells. It contains most of the cell's genetic material (DNA), organized as multiple long linear DNA molecules in complex with a large variety of proteins, such as histones, to form chromosomes. The genes within these chromosomes are the cell's nuclear genome. The function of the nucleus is to maintain the integrity of these genes and to control the activities of the cell by regulating gene expression – the nucleus is therefore the control center of the cell. The main structures making up the nucleus are the nuclear envelope, a double membrane that encloses the entire organelle and separates its contents from the cellular cytoplasm, and the nuclear lamina, a meshwork within the nucleus that adds mechanical support, much like the cytoskeleton supports the cell as a whole. Because the nuclear membrane is impermeable to most molecules, nuclear pores are required to allow movement of molecules, including RNA molecules, across the envelope. These pores cross both of the membranes, providing a channel that allows free movement of small molecules and ions. The movement of larger molecules such as proteins is carefully controlled, and requires active transport regulated by carrier proteins. Nuclear transport is crucial to cell function, as movement through the pores is required for both gene expression and chromosomal maintenance. Although the interior of the nucleus does not contain any membrane-bound subcompartments, its contents are not uniform, and a number of

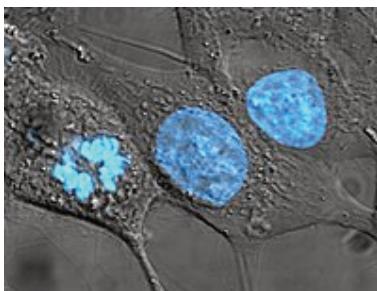


Fig. 1 HeLa cells stained for DNA with the Blue Hoechst dye. The central and rightmost cell are in interphase, thus their entire nuclei are labeled. On the left a cell is going through mitosis and its DNA has condensed ready for division.

subnuclear bodies exist, made up of unique proteins, RNA molecules, and particular parts of the chromosomes. The best known of these is the nucleolus, which is mainly involved in the assembly of ribosomes. After being produced in the nucleolus, ribosomes are exported to the cytoplasm where they translate mRNA.

Bacteria have no nucleus

Bacteria are single-cell organisms that do not have a nucleus and rarely contain any other internal regions separated by membranes. The DNA of bacteria is therefore not packed into a nucleus and is instead openly accessible within the cell. This difference between the structure of bacteria and eukaryotes is so great that it is sometimes considered to be the most important distinction among groups of organisms. Cells lacking a nucleus came to be called prokaryotes. This means that the genomes of prokaryotes are not localized to a compartment as they are localized to the nucleus in eukaryotic cells. In some cases an irregular structure called the nucleoid consists of the DNA of the prokaryotic cell in combination with proteins in an arrangement substantially different from the DNA/protein chromosomes of eukaryotic cells. Another difference between eukaryotes and prokaryotes is in the size of the ribosomes of the cells. Ribosomes, are the structures responsible for the manufacture of proteins. They are therefore key participants in the flow of genetic information from genes to protein. In the 1960's and 1970's, powerful forms of microscopy proved that ribosomes of eukaryotic cells are larger in size than ribosomes in prokaryotic cells. Prokaryotic ribosomes are around 20 nanometers in diameter and are composed of 65% ribosomal RNA (abbreviated rRNA) and 35% protein molecules. Eukaryotic ribosomes are between 25 and 30 nanometers with an rRNA-to-protein ratio that is close to 1.

Molecular biology experiments in the 1990s further clarified some

of these distinctions. By analyzing the sequences of the rRNA molecules in ribosomes from prokaryotic cells it became clear that prokaryotes consist of two very different groups of organisms that evolved from an ancient common ancestor. These two groups, or domains, are called *Bacteria* and *Archaea*. Archaeal cells, though superficially resembling bacteria, have enough unique properties that there presently stands a strong scientific consensus that living cells fall into three domains of life: the *Archaea*, the *Bacteria*, and the *Eukarya* (the single-cell and multicellular organisms that have a nucleus). While ongoing molecular biology studies are working out the details, it is widely agreed that the genome of each present day organism has descended from one of those three main ancestral genomes. Countless rounds of mutation (changes in genomes) and selection (selection of surviving individuals and passing forward of the surviving genome) accounts for the diversity of life. Molecular biology gives us the tools to study a genome's heritage and to evaluate the assorted markers it may share with close or distant kin. Genomes have been going forward over time immemorial. Molecular biology is enabling us to piece together the resulting relationships.

Extrachromosomal elements

In general, prokaryotes lack the following membrane-bound cell compartments: mitochondria and chloroplasts. Instead, processes such as oxidative phosphorylation and photosynthesis take place across the prokaryotic plasma membrane. However, prokaryotes do possess some internal structures, such as cytoskeletons, and the bacterial order Planctomycetes have a membrane around their nucleoid and contain other membrane-bound cellular structures. Both eukaryotes and prokaryotes contain large RNA/protein structures called ribosomes, which produce protein. Prokaryotes are usually much smaller than eukaryotic cells. Prokaryotes also

differ from eukaryotes in that they contain only a single loop of stable chromosomal DNA stored in an area named the nucleoid, whereas eukaryote DNA is found on tightly bound and organized chromosomes. Although some eukaryotes have satellite DNA structures called plasmids, in general these are regarded as a prokaryote feature, and many important genes in prokaryotes are stored on plasmids. Prokaryotes have a larger surface-area-to-volume ratio giving them a higher metabolic rate, a higher growth rate, and, as a consequence, a shorter generation time compared to Eukaryotes. A criticism of this classification is that the word “prokaryote” is based on what these organisms are not (they are not eukaryotic), rather than what they are (either archaea or bacteria). In 1977, Carl Woese proposed dividing prokaryotes into the Bacteria and Archaea (originally Eubacteria and Archaebacteria) because of the major differences in the structure and genetics between the two groups of organisms. This arrangement of Eukaryota (also called “Eukarya”), Bacteria, and Archaea is called the three-domain system, replacing the traditional two-empire system.

What about viruses?

Viruses, which require cells to replicate, are not themselves considered to be living things though they operate and multiply according to many of the same principles found in living cells. For some viruses, the genetic information of the virus is carried by a DNA molecule. These are the DNA viruses. Other viruses carry their genetic information as an RNA molecule, making these the RNA viruses. Some of the viruses whose genomes have been well-studied will be discussed below.

2. Genomes

This chapter is a derivative of New World Encyclopedia Genome (<https://www.newworldencyclopedia.org/entry/Genome>) under a Creative Commons Attribution/Share-Alike License 3.0 (Unported) and New World Encyclopedia Human Genome (https://www.newworldencyclopedia.org/entry/Human_genome), licensed under Creative Commons Attribution/ Share-Alike License 3.0 (Unported).

Genomes

In general, a genome is one complete set of hereditary information that characterizes an organism, as encoded in the DNA (or, in the case of some viruses, RNA). That is, a genome is equivalent to the complete genetic sequence on one of the two sets of chromosomes of the somatic cells of a diploid individual, or the total genetic sequence in the single chromosome of a bacteria, or the sequence of RNA in an RNA virus. The genome includes both the genes and the non-coding sequences of DNA.

In eukaryotes, the term genome can be applied specifically to mean that genetic content 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.

The sequencing and comparison of the genomes of diverse organisms shows the remarkable connectedness of living organisms, as even more complex species, higher on the phylogenetic tree, share basic sequences with bacteria. Many sequences in the genomes of the yeast *Saccharomyces*, the fruit fly *Drosophila*, and the worm *Caenorhabditis* are the same, coding for the same genes.

The complexity of the genome is also evident. An analogy to the haploid human genome stored on DNA is that of data stored on a hard drive:

- The hard drive contains 3.3 billion bytes of information (3.3 gigabytes);
- The information on the hard drive is distributed into 23 folders of varying size and content. On average, each folder contains 1/23 of the 3.3 billion bytes on the hard drive, or 140 million bytes (140 megabytes).
- All but one of the folders is named by a number ranging from 1 to 22 in approximate descending order of folder size. The exception is the “sex-determining folder” which is named as either X or Y depending on which was included on the hard drive.
- The hard drive plugs into the cell nucleus of a germ cell, occupying a volume no bigger than a
- A copy of the book (all 5000 volumes) is contained in almost every cell.

The units of heredity in living organisms are encoded in an organism's genetic material, DNA. The nucleic acid DNA (deoxyribonucleic acid) contains the genetic instructions used in the development and functioning of all known living organisms. (Some viruses utilize RNA, but are not universally considered living organisms.) The main role of DNA molecules is the long-term storage of information. DNA teams with the nucleic acid RNA (ribonucleic acid) to together oversee and carry out the construction of the tens of thousands of protein molecules needed by living organisms.

As nucleic acids, DNA and RNA contain numerous nucleotides (each composed of a phosphate unit, a sugar unit, and a “base” unit) linked recursively through the sugar and phosphate units to form a long chain with base units protruding from it. Nucleic acids carry the coded genetic information of life according to the order of the

base units extending along the length of the molecule. The DNA, which carries genetic information in cells, is normally packaged in the form of one or more large macromolecules called chromosomes.

Genome refers to the total DNA sequence that characterizes a species. That it, a genome is the genetic content (DNA sequences) contained within one set of chromosomes in eukaryotes, or the single chromosome of prokaryotes. For those viruses that utilize only RNA as hereditary material, genome is equivalent to the RNA sequence. Genome includes not only the coding genes of a chromosome but also the non-coding sequences, sometimes referred to as “junk DNA.” In humans, this non-coding DNA may be as much as 97% of the total DNA.

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.

The size of genomes is measured in terms of the number of base pairs, although the large numbers mean that the unit used tends to be *megabases* (Mb), corresponding to 1,000 base pairs.

Genomes of organelles

Most biological entities that are more complex than a virus sometimes or always carry additional genetic material besides that

which resides in their chromosomes. The plasmids of plants and algae, such as chloroplasts, carry genetic material within their membranes, separate and distinct from that of the nucleus. Likewise, the mitochondria of all eukaryotes contain genetic material within their membranes as well, separate and distinct from the nuclear DNA.

Generally, in eukaryotes such as plants, protozoa, and animals, the term “genome” carries the typical connotation of only information on chromosomal DNA. So although these organisms contain mitochondria that have their own DNA, the genes in this mitochondrial DNA are not considered part of the genome. Instead, mitochondria or chloroplasts are sometimes said to have their own genome, often referred to as the “mitochondrial genome” or chloroplast genome.

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

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.

Genome determinations and species comparisons

Technology has developed whereby it is possible to determine the entire DNA sequence of an organism's genome. 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 5368 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. Genomes were subsequently elucidated for several bacteria (including *Escherichia coli*), then yeast (*Saccharomyces*), a plant (*Arabidopsis*), and some animals (the nematode *Caenorhabditis* and the fruit fly *Drosophila*).

The genome of numerous organisms has since been done. The Human Genome Project was organized to map and to sequence the human genome. The completion of the essential sequence of the human genome was announced in June 2000. Other genome projects include mouse, rice, and so forth, with the cost of sequencing continuing to drop and making the process more feasible. In May 2007, the full genome of DNA pioneer James D. Watson was recorded, perhaps a gateway to upcoming personalized genomic medicine.

One of the more interesting results from comparing the genomes of various organisms is that there are basic genes of higher organism that can be traced back to genes in bacteria

In general, genome size is larger for organisms higher on the phylogenetic tree, with humans having a genome of about 3500 Mb and a bacterium only about 4 Mb. However, the presence of coding and non-coding DNA also is reflected in many organisms, such as lungfishes and salamanders, having unusually large genomes.

Genomes of various species

Organism	Genome size (base pairs)	Note
Virus, Bacteriophage MS2	3,569	First sequenced RNA-genome
Virus, SV40		
Virus, Phage Φ-X174;	5,386	First sequenced DNA-genome
Virus, Phage λ	50,000	
Bacterium, <i>Haemophilus influenzae</i>	1,830,000	First genome of living organism, July 1995
Bacterium, <i>Carsonella ruddii</i>	160,000	Smallest non-viral genome.
Bacterium, <i>Buchnera aphidicola</i>	600,000	
Bacterium, <i>Wigglesworthia glossinidia</i>	700,000	
Bacterium, <i>Escherichia coli</i>	4,000,000	
Plant, <i>Arabidopsis thaliana</i>	157,000,000	First plant genome sequenced, Dec 2000.
Plant, <i>Genlisea margaretae</i>	63,400,000	Smallest recorded flowering plant genome, 2006
Plant, <i>Fritillaria assyrica</i>	130,000,000,000	
Plant, <i>Populus trichocarpa</i>	480,000,000	First tree genome, Sept 2006
Yeast, <i>Saccharomyces cerevisiae</i>	20,000,000	
Fungus, <i>Aspergillus nidulans</i>	30,000,000	
Nematode, <i>Caenorhabditis elegans</i>	98,000,000	First multicellular animal genome, December 1998

Insect, <i>Drosophila melanogaster</i> aka Fruit Fly	130,000,000	
Insect, <i>Bombyx mori</i> aka Silk Moth	530,000,000	
Insect, <i>Apis mellifera</i> aka Honeybee	1,770,000,000	
Fish, <i>Tetraodon nigroviridis</i> , type of Puffer fish	385,000,000	Smallest vertebrate genome known
Mammal, <i>Homo sapiens</i>	3,200,000,000	
Fish, <i>Protopterus aethiopicus</i> aka Marbled lungfish	130,000,000,000	Largest vertebrate genome known

Note: The DNA from a single human cell has a length of ~1.8 meters (but at 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

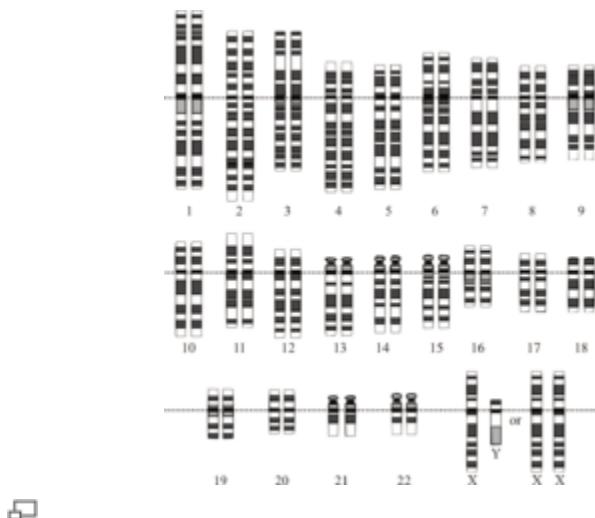
Genomes are more than the sum of an organism's genes and have traits that may be measured and studied without reference to the details of any particular genes and their products. Researchers compare traits such as *chromosome number* (karyotype), genome size, gene order, codon usage bias, and GC-content 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 acquisition of entire sets of genes, even whole genomes of organisms, has been postulated as a major source of transmitted variation in organisms. And eukaryotic cells seem to have experienced a transfer of some genetic material from their chloroplast and mitochondrial genomes to their nuclear chromosomes.

Human genome



A graphical representation of the normal human karyotype.

The human genome is the genome of *Homo sapiens*; that is, the hereditary information that genetically characterizes human beings as encoded on the DNA of one set of the 23 chromosome pairs of the somatic cells. Twenty-two of these are autosomal chromosome pairs, while the remaining pair is sex-determining. As the complete genetic sequence of one of the two sets of chromosomes, the human genome includes both the genes and the non-coding sequences of DNA.

The Human Genome Project produced a reference sequence of the human genome, which is used worldwide in biomedical sciences. The haploid human genome occupies a total of just over 3 billion DNA base pairs and has a data size of approximately 750 megabytes (Overbye 2007). This haploid human genome contains an estimated 20,000 to 25,000 protein-coding genes, far fewer than had been expected before its sequencing (IHGSC 2004). In fact, only about 1.5 percent of the genome codes for proteins, while the rest is comprised of RNA genes, regulatory sequences, introns, and (controversially) “junk” DNA (IHGSC 2001).

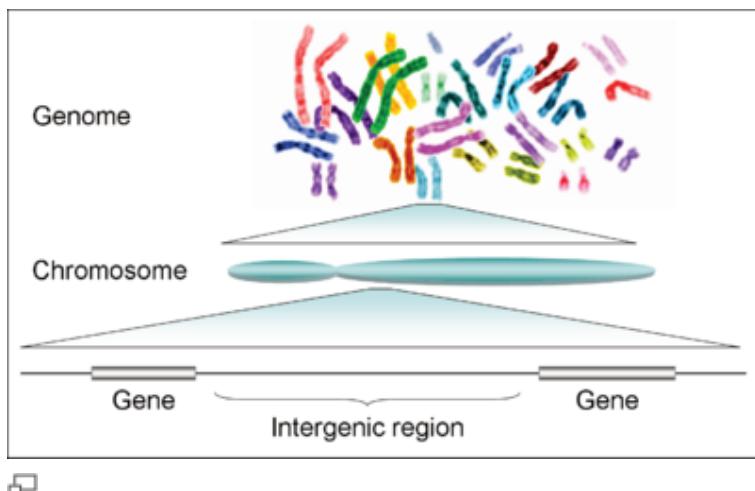
The tremendous breakthrough in resolving the genomes of many species, including humans, has been of great value in understanding organisms and their connectedness over time. However, this does not imply that mapping every gene that makes up a person will allow one to explain that person. In addition to the importance of environmental factors, various religious perspectives hold that life cannot be explained by physico-chemical processes alone and that human beings are more than just physical beings, possessing also a spiritual aspect.

Understanding the human genome is helpful in understanding and working toward a resolution of genetic diseases. Some attention also must be given to lifestyle choices and environmental factors, since they can contribute to genetic damage within one's own cells, such as through exposure to harmful chemicals or radiation, drug use, or infection with a pathogen. Recently, an active area of research has been epigenetics, including to what extent DNA can

be modified or imprinted by one's experiences, such as via diet, smoking, or obesity (Leake 2008).

Features

Chromosomes



The human genome is composed of 23 pairs of chromosomes (46 in total), each of which contain hundreds of genes separated by *intergenic regions*. Intergenic regions may contain regulatory sequences and non-coding DNA.

There are 24 distinct human chromosomes: 22 autosomal chromosomes, plus the sex-determining X and Y chromosomes. Chromosomes 1–22 are numbered roughly in order of decreasing size. Somatic cells usually have 23 chromosome pairs: One copy of chromosomes 1–22 from each parent, plus an X chromosome from

the mother, and either an X or Y chromosome from the father, for a total of 46 chromosomes.

Genes

There are estimated 20,000 to 25,000 human protein-coding genes. The estimate of the number of human genes has been repeatedly revised down from initial predictions of 100,000 or more as genome sequence quality and gene finding methods have improved, and could continue to drop further.

Surprisingly, the number of human genes seems to be less than a factor of two greater than that of many much simpler organisms, such as the roundworm and the fruit fly. However, human cells make extensive use of alternative splicing to produce several different proteins from a single gene, and the human proteome (entire complement of proteins expressed by a genome) is thought to be much larger than those of the aforementioned organisms. In addition, most human genes have multiple exons, and human introns are frequently much longer than the flanking exons.

Human genes are distributed unevenly across the chromosomes. Each chromosome contains various gene-rich and gene-poor regions, which seem to be correlated with chromosome bands and GC-content. The significance of these nonrandom patterns of gene density is not well understood. In addition to protein coding genes, the human genome contains thousands of RNA genes, including tRNA, ribosomal RNA, microRNA, and other non-coding RNA genes.

Regulatory sequences

The human genome has many different regulatory sequences that are crucial to controlling gene expression. These are typically short

sequences that appear near or within genes. A systematic understanding of these regulatory sequences and how they together act as a gene regulatory network is only beginning to emerge from computational, high-throughput expression and comparative genomics studies.

Identification of regulatory sequences relies in part on the concept of evolutionary conservation. The evolutionary branch between the human and mouse, for example, is considered to have occurred 70 to 90 million years ago . So computer comparisons of gene sequences that identify conserved non-coding sequences will be an indication of their importance in duties such as gene regulation .

Another comparative genomic approach to locating regulatory sequences in humans is the gene sequencing of the puffer fish. These vertebrates have essentially the same genes and regulatory gene sequences as humans, but with only one-eighth the “junk” DNA. The compact DNA sequence of the puffer fish makes it much easier to locate the regulatory genes .

Other DNA

Protein-coding sequences (specifically, coding exons) comprise less than 1.5 percent of the human genome. Aside from genes and known regulatory sequences, the human genome contains vast regions of DNA the function of which, if any, remains unknown. These regions in fact comprise the vast majority, by some estimates 97 percent, of the human genome size. Much of this is composed of:

Repeat elements	Transposons	Pseudogenes
<ul style="list-style-type: none"> • Tandem repeats <ul style="list-style-type: none"> ◦ Satellite DNA ◦ Minisatellite ◦ Microsatellite • Interspersed repeats <ul style="list-style-type: none"> ◦ SINEs ◦ LINEs 	<ul style="list-style-type: none"> • Retrotransposons <ul style="list-style-type: none"> ◦ LTR <ul style="list-style-type: none"> ▪ Ty1-copia ▪ Ty3-gypsy ◦ Non-LTR <ul style="list-style-type: none"> ▪ SINEs ▪ LINEs • DNA Transposons 	

However, there is also a large amount of sequence that does not fall under any known classification.

Much of this sequence may be an evolutionary artifact that serves no present-day purpose, and these regions are sometimes collectively referred to as “junk” DNA. There are, however, a variety of emerging indications that many sequences within likely are functional but in ways that are not fully understood. Recent experiments using microarrays have revealed that a substantial fraction of non-genic DNA is, in fact, transcribed into RNA, which leads to the possibility that the resulting transcripts may have some unknown function. Also, the evolutionary conservation across the mammalian genomes of much more sequence than can be explained by protein-coding regions indicates that many, and perhaps most, functional elements in the genome remain unknown. The investigation of the vast quantity of sequence information in the human genome whose function remains unknown is currently a major avenue of scientific inquiry.

Mitochondrial genome

The mitochondria of human beings also contain genetic material within their membranes, separate and distinct from the nuclear DNA. Generally, the term “human genome” carries the connotation

of only information on chromosomal DNA. Thus, the genes in the mitochondrial DNA are not considered part of the human genome, although such may be referred to as the “mitochondrial genome.”

The human mitochondrial genome, while usually not included when referring to the “human genome,” is of tremendous interest to geneticists, since it undoubtedly plays a role in mitochondrial disease. It also sheds light on human evolution; for example, analysis of variation in the human mitochondrial genome has led to the postulation of a recent common ancestor for all humans on the maternal line of descent.

Due to the lack of a system for checking for copying errors, Mitochondrial DNA (mtDNA) has a more rapid rate of variation than nuclear DNA. This 20-fold increase in the mutation rate allows mtDNA to be used for more accurate tracing of maternal ancestry. Studies of mtDNA in populations have allowed ancient migration paths to be traced, such as the migration of Native Americans from Siberia or Polynesians from southeastern Asia. It has also been used to show that there is no trace of Neanderthal DNA in the European gene mixture inherited through purely maternal lineage.

Variation

Most studies of human genetic variation have focused on single nucleotide polymorphisms (SNPs), which are substitutions in individual bases along a chromosome. Most analyses estimate that SNPs occur on average somewhere between every 1 in 100 and 1 in 1,000 base pairs in the euchromatic human genome, although they do not occur at a uniform density. Thus follows the popular statement that “we are all, regardless of race, genetically 99.9 percent the same”, although this would be somewhat qualified by most geneticists. For example, a much larger fraction of the genome is now thought to be involved in copy number variation. A large-

scale collaborative effort to catalog SNP variations in the human genome has been undertaken.

The genomic loci and length of certain types of small repetitive sequences are highly variable from person to person, which is the basis of DNA fingerprinting and DNA paternity testing technologies. The heterochromatic portions of the human genome, which total several hundred million base pairs, are also thought to be quite variable within the human population (they are so repetitive and so long that they cannot be accurately sequenced with current technology). These regions contain few genes, and it is unclear whether any significant phenotypic effect results from typical variation in repeats or heterochromatin.

Genetic disorders

Most gross genomic mutations in germ cells probably result in inviable embryos; however, a number of human diseases are related to large-scale genomic abnormalities. Down syndrome, Turner Syndrome, and a number of other diseases result from nondisjunction of entire chromosomes. Cancer cells frequently have aneuploidy of chromosomes and chromosome arms, although a cause and effect relationship between aneuploidy and cancer has not been established.

Most aspects of human biology involve both genetic (inherited) and non-genetic (environmental) factors. Some inherited variation influences aspects of our biology that are not medical in nature (height, eye color, ability to taste or smell certain compounds, and so on). Moreover, some genetic disorders only cause disease in combination with the appropriate environmental factors (such as diet).

With these caveats, genetic disorders may be described as clinically defined diseases caused by genomic DNA sequence variation. In the most straightforward cases, the disorder can be

associated with variation in a single gene. For example, cystic fibrosis is caused by mutations in the CFTR gene, and is the most common recessive disorder in Caucasian populations with over 1300 different mutations known. Disease-causing mutations in specific genes are usually severe in terms of gene function, and are fortunately rare, thus genetic disorders are similarly individually rare. However, since there are many genes that can vary to cause genetic disorders, in aggregate they comprise a significant component of known medical conditions, especially in pediatric medicine. Molecularly characterized genetic disorders are those for which the underlying causal gene has been identified; currently about 2200 such disorders have been annotated.

Studies of genetic disorders are often performed by means of family-based studies. In some instances, population based approaches are employed, particularly in the case of so-called founder populations such as those in Finland, French-Canada, Utah, Sardinia, and so forth. Diagnosis and treatment of genetic disorders are usually performed by a geneticist-physician trained in clinical/medical genetics. The results of the Human Genome Project are likely to provide increased availability of genetic testing for gene-related disorders, and eventually improved treatment. Parents can be screened for hereditary conditions and counseled on the consequences, the probability it will be inherited, and how to avoid or ameliorate it in their offspring.

As noted above, there are many different kinds of DNA sequence variation, ranging from complete extra or missing chromosomes down to single nucleotide changes. It is generally presumed that much naturally occurring genetic variation in human populations is phenotypically neutral, that is, has little or no detectable effect on the physiology of the individual. Genetic disorders can be caused by any or all known types of sequence variation. To molecularly characterize a new genetic disorder, it is necessary to establish a causal link between a particular genomic sequence variant and the clinical disease under investigation. Such studies constitute the realm of human molecular genetics.

With the advent of the Human Genome Project, it has become feasible to explore subtle genetic influences on many common disease conditions such as diabetes, asthma, migraine, schizophrenia, and so forth. Although some causal links have been made between genomic sequence variants in particular genes and some of these diseases, often with much publicity in the general media, these are usually not considered to be genetic disorders per se, as their causes are complex, involving many different genetic and environmental factors. Thus, there may be disagreement in particular cases whether a specific medical condition should be termed a genetic disorder.

Of course, as beings that are not just physical, but also are mental, social, and spiritual in nature, many factors interplay with genetic disorders, not just physical factors. A person who leads an unhealthy life, physically or spiritually, either by choice or ignorance, can contribute to the genetic damage within his or her own cells. Damage to germ cells can be passed down to one's descendants in the form of mutations or chromosomal disorders. For example, a person may be exposed to harmful chemicals or radiation, perhaps as a result of warfare or careless disposal of radioactive materials (environmental pollution). A person may engage in careless or promiscuous sex and become infected with a pathogen that can lead to genetic damage. Drug use is another correlate of genetic damage. Sometimes a person may act conscientiously, yet be infected because of societal failure. An example of this is the use of thalidomide, a prescribed drug that later was found to cause birth defects when taken during pregnancy.

Similarly, a person's actions can impact the expression of certain genetic disorders. For example, phenylketonuria (PKU) is a genetic disorder characterized by a deficiency in the enzyme phenylalanine hydroxylase (PAH), which is necessary to metabolize the amino acid phenylalanine to tyrosine. However, PKU can be controlled by diet. A diet low in phenylalanine and high in tyrosine can bring about a nearly total cure.

Evolution

Comparative genomics studies of mammalian genomes suggest that approximately 5 percent of the human genome has been conserved by evolution since the divergence of those species approximately 200 million years ago, containing the vast majority of genes. Intriguingly, since genes and known regulatory sequences probably comprise less than 2 percent of the genome, this suggests that there may be more unknown functional sequence than known functional sequence.

A smaller, yet large, fraction of human genes seem to be shared among most known vertebrates. The chimpanzee genome is 95 percent identical to the human genome. On average, a typical human protein-coding gene differs from its chimpanzee ortholog by only two amino acid substitutions; nearly one third of human genes have exactly the same protein translation as their chimpanzee orthologs. A major difference between the two genomes is human chromosome 2, which is equivalent to a fusion product of chimpanzee chromosomes 12 and 13.

3. DNA and RNA molecules

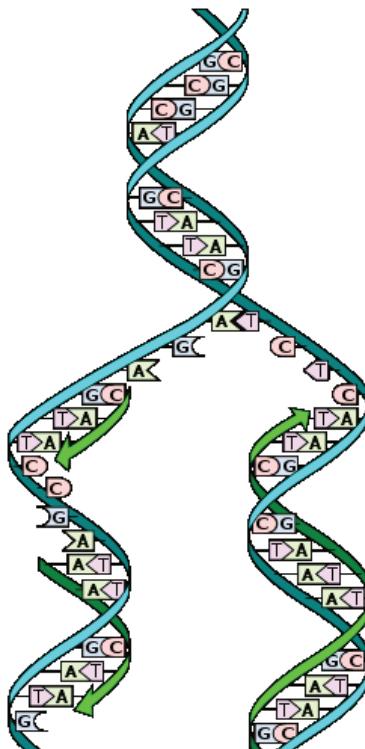
The information-rich structure of DNA

The genetic blueprint for living cells resides in the long nucleic acid molecule known as deoxyribonucleic acid, or DNA. DNA is made of simple units that line up in a particular order within this large molecule. The order of these units carries genetic information, similar to how the order of letters on a page carry information. The language used by DNA is called the genetic code, which lets organisms read the information in the genes. This information is the instructions for constructing and operating a living organism. 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). DNA is copied and inherited across generations, providing the continuity between ancestors and descendants.

DNA molecules are ideally structured for 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.

The structure of DNA was first discovered by James D. Watson and Francis Crick. It is the same for all species, comprising two helical chains each coiled round the same axis, each with a pitch of 34 Ångströms (3.4 nanometres) and a radius of 10 Ångströms (1.0 nanometres). 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. 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.

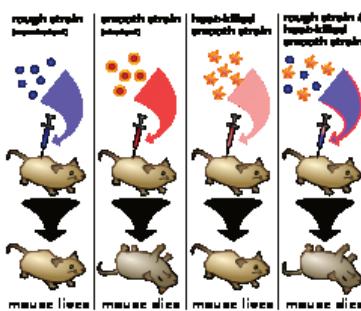


DNA replication. DNA is unwound and nucleotides are matched to make two new strands.

Early proof that DNA is the transforming principle

In 1928, Frederick Griffith conducted one of the first experiments suggesting that bacteria are capable of transferring genetic information through a process known as transformation.

Griffith used two strains of *Streptococcus pneumoniae* bacteria which infect mice – a type III-S (smooth) and type II-R (rough) strain. The III-S strain covers itself with a polysaccharide capsule that protects it from the host's immune system, resulting in the death of the host, while the II-R strain doesn't have that protective capsule and is defeated by the host's immune system. A German bacteriologist, Fred Neufeld, had discovered the three pneumococcal types (Types I, II, and III) and discovered the Quellung reaction to identify them in vitro. Until Griffith's experiment, bacteriologists believed that the types were fixed and unchangeable, from one generation to another. In this experiment, bacteria from the III-S strain were killed by heat, and their remains were added to II-R strain bacteria. While neither alone harmed the mice, the combination was able to kill its host. Griffith was also able to isolate both live II-R and live III-S strains of pneumococcus from the blood of these dead mice. Griffith concluded that the type II-R had been “transformed” into the lethal III-S strain by a “transforming principle” that was somehow part of the dead III-S strain bacteria.



Griffith's experiment discovering the “transforming principle” in *pneumococcus* bacteria.

Today, we understand that the “transforming principle” Griffith observed was the DNA of the III-S strain bacteria. While the bacteria had been killed, the DNA had survived the heating process and was taken up by the II-R strain bacteria. The III-S strain DNA contains the genes that form the protective polysaccharide capsule. Equipped with this gene, the former II-R strain bacteria were now protected from the host’s immune system and could kill the host. The exact nature of the transforming principle (DNA) was verified in the experiments done by Avery, MacLeod and McCarty and by Hershey and Chase.

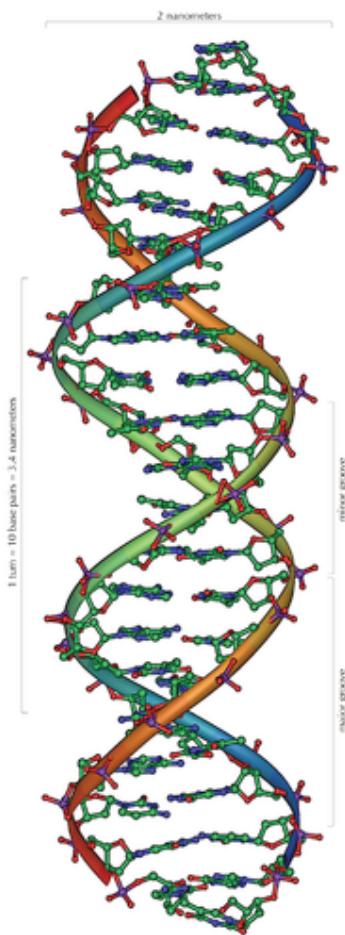
First confirmation:

Alfred Hershey and Martha Chase conducted series of experiments in 1952 by , confirming that DNA was the genetic material, which had first been demonstrated in the 1944 Avery–MacLeod–McCarty experiment. These experiments are known as Hershey Chase experiments. The existence of DNA was known to biologists since 1869, most of them assumed that proteins carried the information for inheritance that time. Hershey and Chase conducted their experiments on the T2 phage. The phage consists of a protein shell containing its genetic material. The phage infects a bacterium by attaching to its outer membrane and injecting its genetic material and leaving its empty shell attached to the bacterium. In their first set of experiments, Hershey and Chase labeled the DNA of phages with radioactive Phosphorus-32 (p32) (the element phosphorus is present in DNA but not present in any of the 20 amino acids which are component of proteins). They allowed the phages to infect E. coli, and through several elegant experiments were able to observe the transfer of P32 labeled phage DNA into the cytoplasm of the bacterium. In their second set of experiments, they labeled the phages with radioactive Sulfur-35 (Sulfur is present in the amino acids cysteine and methionine, but not in DNA). Following infection of E. coli they then sheared the viral protein shells off of infected cells using a high-speed blender and separated the cells and viral coats by using a centrifuge. After separation, the radioactive S35 tracer was observed in the protein

shells, but not in the infected bacteria, supporting the hypothesis that the genetic material which infects the bacteria was DNA and not protein.

The double helix

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



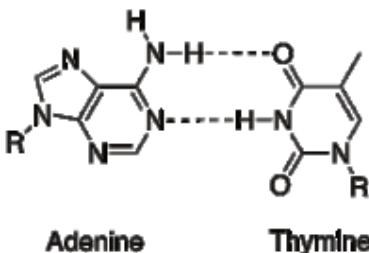
Structure of DNA.

Base pairing Of DNA

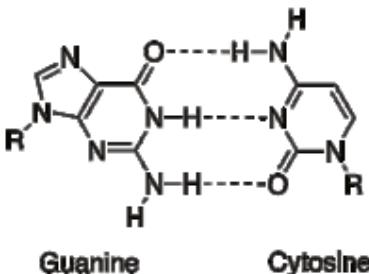
In molecular biology, two nucleotides on opposite complementary DNA strands that are connected via hydrogen bonds are called a base pair (often abbreviated *bp*).

In the canonical Watson-Crick DNA base pairing, Adenine (A) forms a base pair with Thymine (T) and Guanine (G) forms a base pair with Cytosine (C). In (U). Alternate hydrogen bonding pair and Hoogsteen base RNA-giving rise to complex and

"Chargaff's rules", which had been a mystery, state that DNA from any cell or organism has a content of guanine equal to cytosine, and a content of adenine equal to thymine. Base pairing of complementary DNA strands provides the explanation



An AT base pair demonstrating two intermolecular hydrogen bonds.



A **GC** base pair demonstrating three intermolecular hydrogen bonds.

Example

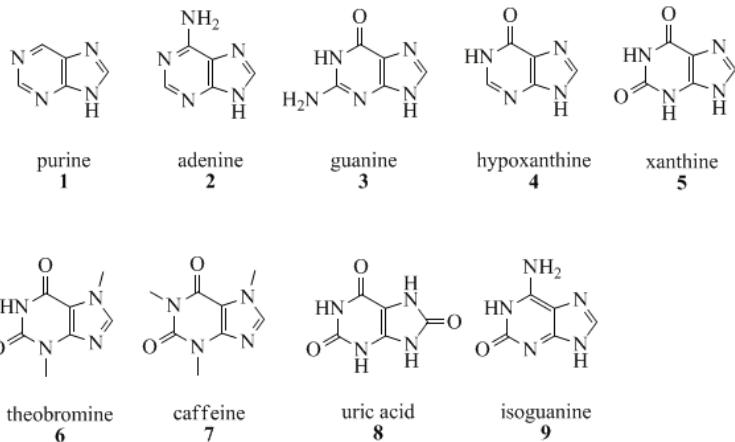
5' CTCGTTGCGCTCTATCG3'
3' GAGCAAACGCGAGATAGC5'

Purines and pyrimidines

Purines

The German chemist Emil Fischer in 1884 gave the name ‘purine’ to molecules that have two fused rings in the arrangement shown by structure 1 below (Fischer synthesized purine for the first time in 1899 from uric acid which had been isolated from kidney stones). The adenine (A) and guanine (G) of DNA and RNA are purines, as are a number of other important biomolecules, including caffeine.

Example purines:



Adenine

Adenine is one of the two purine nucleobases (the other being guanine) used in forming nucleotides of the nucleic acids (DNA or RNA). In DNA, adenine binds to thymine via two hydrogen bonds to assist in stabilizing the nucleic acid structures. Adenine forms adenosine, a nucleoside, when attached to ribose, and

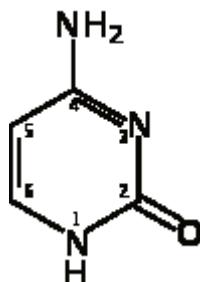
deoxyadenosine when attached to deoxyribose. It forms adenosine triphosphate (ATP), a nucleotide, when three phosphate groups are added to adenosine. When an RNA strand is being synthesized as a biological process, ATP serves as one of the chief building blocks. Similarly, the deoxy form of ATP, known as deoxyadenosine triphosphate (abbreviated deoxyATP or simply dATP), serves as one of the building blocks in the manufacture of DNA strands.

Guanine

Guanine is also a purine. It is present in both DNA and RNA, pairing with cytosine. It binds to cytosine through three hydrogen bonds. Guanine forms guanosine, a nucleoside, when attached to ribose, and deoxyguanosine when attached to deoxyribose. It forms guanosine triphosphate (GTP), a nucleotide, when three phosphate groups are added to guanosine. GTP is used to build RNA whereas dGTP is one of the building blocks for DNA.

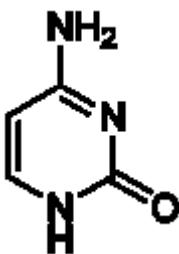
Pyrimidines

The name pyrimidine applies to aromatic organic compound similar to benzene and pyridine, containing two nitrogen atoms at positions 1 and 3 of the six-member ring. Three nucleobases found in nucleic acids, cytosine (C), thymine (T), and uracil (U), are pyrimidines.



Cytosine

Cytosine can be found as part of DNA, as part of RNA, or as a part of a nucleotide. As cytidine triphosphate (CTP), it is the building block of RNA and as deoxyCTP (dCTP) it is the building block of DNA. The nucleoside of cytosine is cytidine. In DNA and RNA, cytosine is paired with guanine. However, it is inherently unstable, and can change into uracil (spontaneous deamination). This can lead to a point mutation if not repaired by DNA repair enzymes. Cytosine can also be methylated into 5-methylcytosine by an enzyme called DNA methyltransferase in a process that has effects on how the information in the DNA molecule is recognized inside the cell.

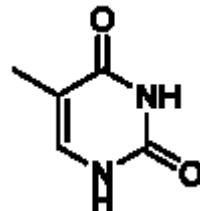


Chemical structure
of cytosine

Cytosine with
numbered
components.
Methylation on
carbon number 5
gives thymine.

Thymine

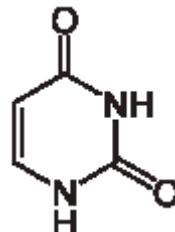
Thymine (T, Thy) is a pyrimidine nucleobase that is one of the four in DNA (A, G, C and T). In RNA, thymine is replaced with uracil in most cases (A, G, C and U). In DNA, thymine(T) binds to adenine (A) via two hydrogen bonds, thus stabilizing the nucleic acid structures.



Chemical structure of thymine

Uracil

Uracil found in RNA, base-pairs with adenine by two hydrogen bonds. Uracil can also form base-pair with any of the bases, depending on how the RNA molecule is arranged. Thus, unlike complementary strands of DNA that seldom deviate from the base-pairing rules, uracil's ability to form nonstandard base-pairs is one of the reasons why RNA structures have more three-dimensional variety than is seen in comparison to the DNA double helix



Chemical structure of uracil

Attachment of nucleobases to a sugar. Nucleosides, and deoxynucleosides

Nucleosides consist of a nucleobase (often referred to as simply base) bound to a ribose or deoxyribose sugar. The five carbon atoms in the sugar ring carbons are numbered from 1

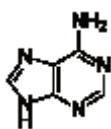
image

The sugar in RNA is ribose. The sugar in DNA is deoxyribose. From ResearchGate.

through 5 with a prime (') mark added to distinguish the sugar's numbering from the numbering of atoms in the nucleic acid. The hydroxy group (OH) on the 1' carbon atom points up, a feature that is highlighted by the greek letter beta (β). The hydroxy group is not present on the 2' carbon in deoxyribose, an important distinction of DNA. The hydroxy groups on the 3' and 5' carbons are the connecting points from one building block to the next in both RNA and DNA molecules.

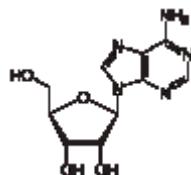
The nucleosides in RNA include adenosine, guanosine, uridine and cytidine. For DNA, the nucleosides are technically termed deoxynucleosides (the term nucleoside is often used with the tacit assumption that if DNA is under discussion the nucleosides are all deoxynucleosides). The four nucleosides in DNA include deoxyadenosine, deoxyguanosine, thymidine (which is also known as deoxythymidine), and deoxycytidine. Nucleosides are linked to one another by phosphates to form chains of RNA or DNA.

Nucleobase

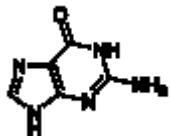


Adenine

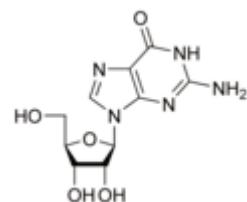
Nucleoside



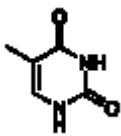
Adenosine
A



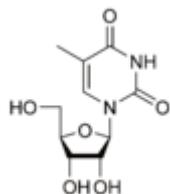
Guanine



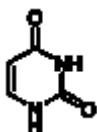
Guanosine
G



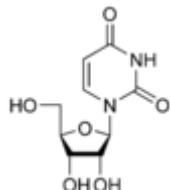
Thymine



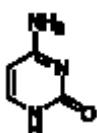
5-Methyluridine
 m^5U



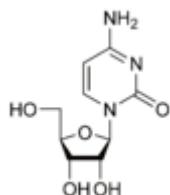
Uracil



Uridine
U



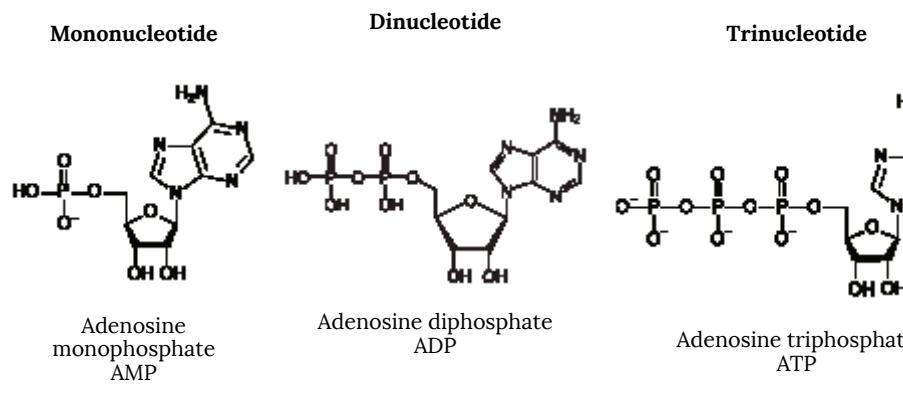
Cytosine



Cytidine
C

Nucleotides

To assemble chains of DNA or RNA, nucleosides are linked together by a single phosphate group. The unit of structure is therefore a *nucleotide*, composed of a nucleoside and a phosphate group. The term mononucleotide is also used to indicate that a nucleic acid chain is essentially a chain of nucleosides bridged by single phosphate groups. However, the process of building a nucleic acid chain requires trinucleotides, which are nucleosides with three phosphates linked to the 5' hydroxyl group of the sugar. Note that the term “building block” is somewhat ambiguous because, on one hand, in the building process, the trinucleotides precursor molecules can be thought of as the building blocks, whereas on the other hand, if a nucleic acid chain is already assembled, the mononucleotides standing in place can be looked upon as the building blocks.



Phosphodiester linkages

When deoxyribonucleotides polymerize to form DNA, the

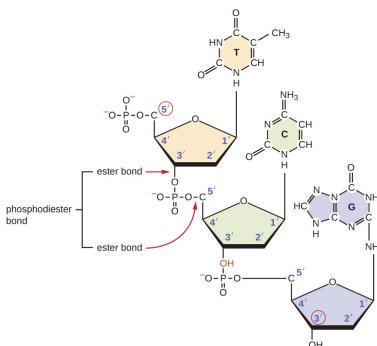
phosphate group from one nucleotide will bond to the 3' carbon on another nucleotide, forming a *phosphodiester linkage*. RNA is linked together similarly.

Chemically speaking, a phosphodiester linkage includes two ester bonds. The first is the ester bond between a phosphoric acid and the 3' hydroxyl of a sugar. The second is the ester bond between that same phosphoric acid and the 5' hydroxyl of the next sugar. Since there are two esters, it is proper to refer to the linkage as a phosphodiester.

Phosphodiester bonds from one sugar to the next supply the backbone to chains of DNA

and RNA. The phosphate groups in the backbone are negatively-charged because of the acidic nature of phosphoric acid (at pH 7 the phosphoric acid loses a proton, leaving behind a negative charge). The backbone of a nucleic acids is therefore negatively charged, and in order for two DNA strands to come together to form a double helix, the repulsive nature of those two negatively charged backbones has to be overcome. Base-pairing (A-T and G-C) between the nucleobases provides much of the attractive force to overcome the repulsion between the backbones. Also helping to neutralize the repulsive force are positively charged metal ions such as magnesium (Mg^{2+}), as well as certain positively charged proteins (such as histone proteins in eukaryotic cells) that bind to the backbone and counteract the negative charges.

In order for the phosphodiester bond to be formed and the nucleotides to be joined, the trinucleotides are broken apart to give off energy, driving the reaction forward. The two phosphates that are broken off remain either remain joined together (forming a



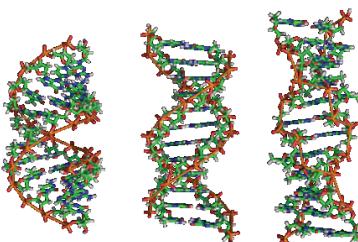
A sequence of three nucleotides joined by two phosphodiester linkages. Beginning from the 5' end, the sequence is TCG. The 5' end is phosphorylated. The 3' end is free (does not have a phosphate attached)

molecule known as pyrophosphate), or are further split into single phosphates, releasing more energy and driving the overall process even harder.

One way of degrading nucleic acids is to cleave the phosphodiester linkages. Hydrolysis, which uses water molecules to split the phosphoester bonds, is commonly catalyzed by digestive enzymes known as nucleases. There are many of these. Some of them are involved in food digestion, and in that role no discrimination is necessary. Other nucleases are highly discriminate, cleaving only certain bonds, usually as a result of sequence recognition by the nuclease. Restriction endonuclease, for example, are enzymes produced by bacteria that cleave only certain sequences. For example, the restriction enzyme known as EcoR1 cleaves double stranded DNA only if it contains the sequence 5'-GAATTC-3'. RNA is a less stable nucleic acid than DNA because the 2' hydroxyl group in the ribose sugars are reactive and if the pH rises too high, the 2' hydroxyls can sever the nearest phosphoester bond, namely the phosphoester bond attached to the 3' hydroxyl.

Forms of DNA

A-DNA: A-DNA is one of the many possible double helical structures of DNA. A-DNA is thought to be one of three biologically active double helical structures along with B- and Z-DNA. It is a right-handed double helix fairly similar to the more common and well-known B-DNA form, but with a shorter more compact helical structure. It appears likely that it occurs only in dehydrated samples of DNA, such as those used in



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

crystallographic experiments, and possibly is also assumed by DNA-RNA hybrid helices and by regions of double-stranded RNA.

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

Z-DNA: Z-DNA is one of the many possible double helical structures of DNA. It is a left-handed double helical structure in which the double helix winds to the left in a zig-zag pattern (instead of to the right, like the more common B-DNA form). Z-DNA is quite different from the right-handed A- and B-forms. The major and minor grooves, unlike A- and B-DNA, show little difference in width. Formation of this structure is generally unfavourable, although certain laboratory conditions can promote it, such as alternating purine-pyrimidine sequence (such as GCGCGCGC ...), or high concentrations of positively charged metal ions. The Z-DNA conformation has been difficult to study because it does not exist as a stable feature of the double helix. Instead, it is a transient structure that is occasionally induced by biological activity and then quickly disappears.

4. The central dogma: Replication and expression of genomes

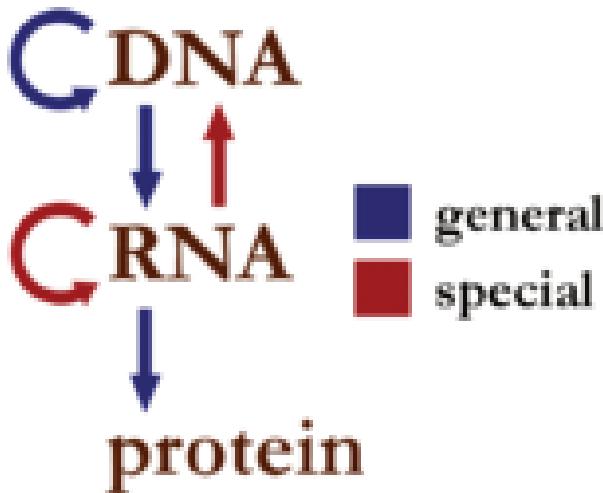
A little history. Who came up with the idea?

The central dogma of molecular biology is an explanation of the flow of genetic information within a biological system. It is often stated as “DNA makes RNA and RNA makes protein,” although this is not its original meaning. It was first stated by Francis Crick in 1957, then published in 1958:

The Central Dogma. This states that once ‘information’ has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to “protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.

— FRANCIS CRICK, 1958

and re-stated in a *Nature* paper published in 1970.



“ The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid. ”

— FRANCIS CRICK

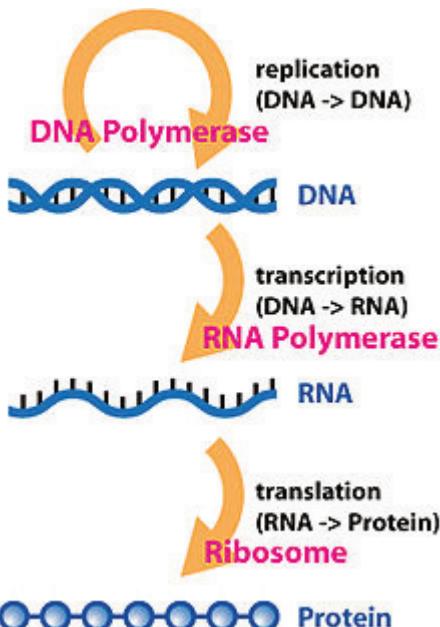
A second version of the central dogma is popular but not precisely correct. This is the simplistic DNA → RNA → protein pathway published by James Watson in the first edition of *The Molecular Biology of the Gene* (1965). Watson's version differs from Crick's because Watson describes a two-step (DNA → RNA and RNA → protein) process as the central dogma. While the dogma, as originally stated by Crick, remains valid today, Watson's version does not.

Why is this idea so central?

The central dogma tells us how the codes of living organisms work. The biopolymers that comprise DNA, RNA and proteins are linear polymers (each monomer is connected to at most two other monomers). The sequence of their monomers effectively encodes information. The transfers of information described by the central dogma ideally are faithful, deterministic transfers, wherein one biopolymer's sequence is used as a template for the construction of another biopolymer with a sequence that is entirely dependent on the original biopolymer's sequence.

The dogma is a framework for understanding the transfer of sequence information between information-carrying biopolymers, in the most common or general case, in living organisms. There are 3 major classes of such biopolymers: DNA and RNA (both nucleic acids), and protein. There are $3 \times 3 = 9$ conceivable direct transfers of information that can occur between these. The dogma classes these into 3 groups of 3: three general transfers (believed to occur normally in most cells), three special transfers (known to occur, but only under specific conditions in case of some viruses or in a laboratory), and three unknown transfers (believed never to occur). The general transfers describe the normal flow of biological information: DNA can be copied to DNA (DNA replication), DNA information can be copied into mRNA (transcription), and proteins can be synthesized using the information in mRNA as a template (translation). The special transfers describe: RNA being copied from RNA (RNA replication), DNA being synthesised using an RNA template (reverse transcription), and proteins being synthesised directly from a DNA template without the use of mRNA. The unknown transfers describe: a protein being copied from a protein, synthesis of RNA using the primary structure of a protein as a template, and DNA synthesis using the primary structure of a protein as a template – these are not thought to naturally occur.

The observed paths of information flow in living cells



The figure to the right sketches out the central dogma of molecular biology. The main directions of information flow in living cells are indicated by arrows (replication, transcription and translation). The main molecular machines involved in those stages are named (DNA polymerase, RNA polymerase, and the ribosome).

Replication

In the sense that DNA replication must occur if genetic material is to be provided for the progeny of any cell, whether somatic or reproductive, the copying from DNA to DNA arguably is the fundamental step in the central dogma. A complex assembly called

the replication fork is where the action takes place in going from the parent strand to the complementary daughter strand.

The replication fork (in bacteria) includes:

- a helicase that unwinds the superhelix as well as the double-stranded DNA helix to create the replication fork
- SSB protein that binds open the double-stranded DNA to prevent it from reassociating
- RNA primase that adds a complementary RNA primer to each template strand as a starting point for replication
- DNA polymerase III that reads the existing template chain from its 3' end to its 5' end and adds new complementary nucleotides from the 5' end to the 3' end of the daughter chain
- DNA polymerase I that removes the RNA primers and replaces them with DNA
- DNA ligase that joins the two Okazaki fragments with phosphodiester bonds to produce a continuous chain

Transcription

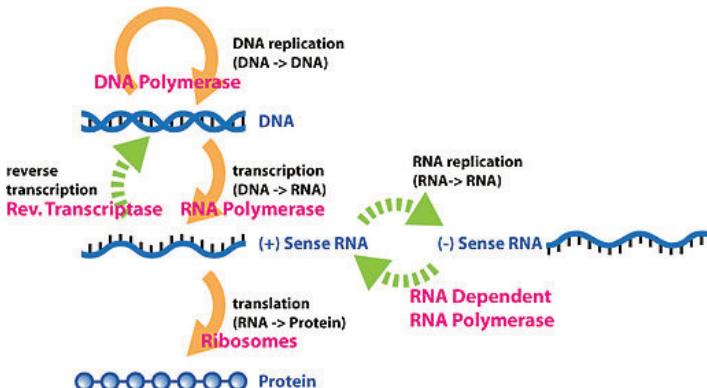
Transcription is the process by which the information contained in a section of DNA is replicated in the form of a newly assembled piece of messenger RNA (mRNA). Enzymes facilitating the process include RNA polymerase and transcription factors. In eukaryotic cells the primary transcript is pre-mRNA. Pre-mRNA must be processed for translation to proceed. Processing includes the addition of a 5' cap and a poly-A tail to the pre-mRNA chain, followed by splicing. Alternative splicing occurs when appropriate, increasing the diversity of the proteins that any single mRNA can produce. The product of the entire transcription process (that began with the production of the pre-mRNA chain) is a mature mRNA chain.

Translation

The mature mRNA finds its way to a ribosome, where it gets translated. In prokaryotic cells, which have no nuclear compartment, the processes of transcription and translation may be linked together without clear separation. In eukaryotic cells, the site of transcription (the cell nucleus) is usually separated from the site of translation (the cytoplasm), so the mRNA must be transported out of the nucleus into the cytoplasm, where it can be bound by ribosomes. The ribosome reads the mRNA triplet codons, usually beginning with an AUG (adenine-uracil-guanine), or initiator methionine codon downstream of the ribosome binding site. Complexes of initiation factors and elongation factors bring aminoacylated transfer RNAs (tRNAs) into the ribosome-mRNA complex, matching the codon in the mRNA to the anti-codon on the tRNA. Each tRNA bears the appropriate amino acid residue to add to the polypeptide chain being synthesised. As the amino acids get linked into the growing peptide chain, the chain begins folding into the correct conformation. Translation ends with a stop codon which may be a UAA, UGA, or UAG triplet.

The mRNA does not contain all the information for specifying the nature of the mature protein. The nascent polypeptide chain released from the ribosome commonly requires additional processing before the final product emerges. For one thing, the correct folding process is complex and vitally important. For most proteins it requires other chaperone proteins to control the form of the product. Some proteins then excise internal segments from their own peptide chains, splicing the free ends that border the gap; in such processes the inside “discarded” sections are called inteins. Other proteins must be split into multiple sections without splicing. Some polypeptide chains need to be cross-linked, and others must be attached to non-protein molecules such as heme before they become functional.

Special information flows



Special transfers of information are highlighted in green in the above figure, including reverse transcription and RNA replication

Reverse transcription

Reverse transcription is the transfer of information from RNA to DNA (the reverse of normal transcription). This is known to occur in the case of retroviruses, such as HIV, as well as in eukaryotes, in the case of retrotransposons and telomere synthesis. It is the process by which genetic information from RNA gets transcribed into new DNA.

RNA replication

RNA replication is the copying of one RNA to another. Many viruses replicate this way. The enzymes that copy RNA to new RNA, called

RNA-dependent RNA polymerases, are also found in many eukaryotes where they are involved in RNA silencing.^[8]

RNA editing, in which an RNA sequence is altered by a complex of proteins and a “guide RNA”, could also be seen as an RNA-to-RNA transfer.

References and online resource

Animation: Central Dogma of Biology

Animation: DNA Replication

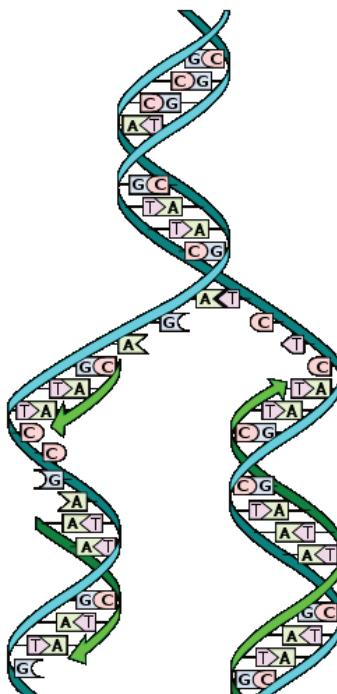
Animation: Transcription of RNA

Animation: mRNA Splicing

Animation: Protein Translation

5. DNA Replication and Repair

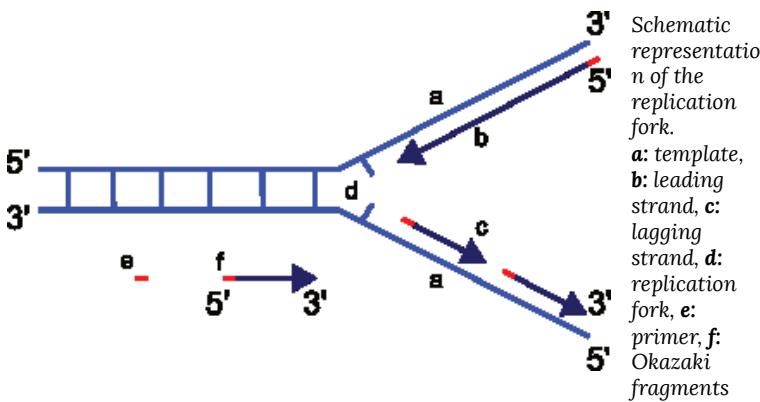
As we know cell division is essential for an organism to grow, but, when a cell divides, it must replicate the DNA (DNA replication take place during S phase) in its genome so that the two daughter cells have the same genetic information as their parent. The double-stranded structure of DNA provides 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.



DNA replication. DNA is unwound and nucleotides are matched to make two new strands.

Replication

In a cell, DNA replication begins at specific locations in the genome, called “**origins**”. Unwinding of DNA at the origin, and synthesis of new strands, forms a **replication fork**. In addition to DNA polymerase, the enzyme that synthesizes the new DNA by adding nucleotides matched to the template strand, a number of other proteins are associated with the fork and assist in the initiation and continuation of DNA synthesis. DNA replication can also be performed *in vitro* (outside a cell). DNA polymerases, isolated from cells, and artificial DNA primers are used to initiate DNA synthesis at known sequences in a template molecule. The polymerase chain reaction (**PCR**), a common laboratory technique, employs such artificial synthesis in a cyclic manner to amplify a specific target DNA fragment from a pool of DNA.



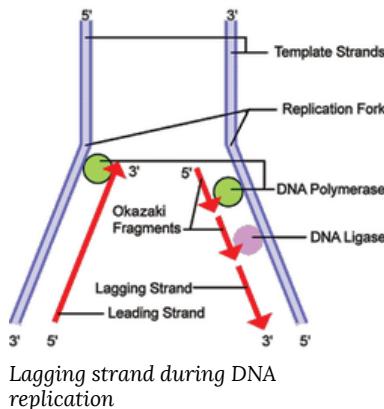
Leading strand

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

a 5'-3' nascent strand. The leading strand is formed along the leading strand template as a polymerase “reads” the template DNA and continuously adds nucleotides to the 3' end of the elongating strand. This polymerase is DNA polymerase III (DNA Pol III) in prokaryotes and presumably Pol ε in eukaryotes.

Lagging strand

The lagging strand template is the coding strand of the DNA double helix that is oriented in a 5' to 3' manner. The newly made lagging strand still is synthesized 5'-3'. However, since the DNA is oriented in a manner that does not allow continual synthesis, only small sections can be read at a time. An RNA primer is placed on the DNA strand 3' to the origin of replication. Just as before, DNA Polymerase reads 3'-5' on the original DNA to produce a 5'-3' nascent strand. Polymerase reaches the origin of replication and stops replication until a new RNA primer is placed 3' to the last RNA primer. These fragments of DNA produced on the lagging strand are called Okazaki fragments. The orientation of the original DNA on the lagging strand prevents continual synthesis. As a result, replication of the lagging strand is more complicated than of the leading strand. On the lagging strand template, primase “reads” the DNA and adds RNA to it in short, separated segments. In eukaryotes, primase is intrinsic to **Pol α**. **DNA polymerase III** or **Pol δ** lengthens the primed segments, forming **Okazaki fragments**. Primer removal in eukaryotes is also performed by Pol δ. In prokaryotes, DNA polymerase I “reads” the



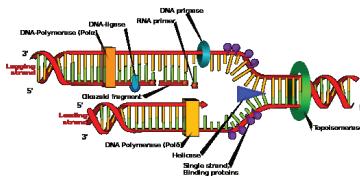
fragments, removes the RNA using its flap endonuclease domain, and replaces the RNA nucleotides with DNA nucleotides (this is necessary because RNA and DNA use slightly different kinds of nucleotides). DNA ligase joins the fragments together.

Okazaki fragment

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

Rate of replication

The rate of DNA replication in a living cell was first measured as the rate of phage T4 DNA elongation in phage-infected *E. coli*. During the period of exponential DNA increase at 37 °C, the rate was 749 nucleotides per second. The mutation rate per base pair per replication during phage T4 DNA synthesis is 1.7 per 10^8 . Thus semiconservative DNA replication is both rapid and accurate.



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 is semiconservative

The **Meselson and Stahl experiment** was an experiment by Matthew Meselson and Franklin Stahl in 1958 which supported the hypothesis that DNA replication was semiconservative.

Semiconservative replication means that when the double stranded DNA helix was replicated, each of the two double stranded DNA helices consisted of one strand coming from the original helix and one newly synthesized. It has been called “the most beautiful experiment in biology.”

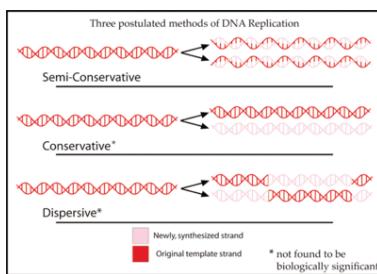
Three hypotheses had been previously proposed for the method of replication of DNA.

In the *semiconservative* hypothesis, proposed by Watson and Crick, the two strands of a parental DNA molecule separate during replication. Each strand then acts as a template for synthesis of a new strand.

The *conservative* hypothesis proposed that the entire parental DNA molecule acted as a template for synthesis of an entirely new one.

The *dispersive* hypothesis is a proposed mechanism that breaks the parental DNA backbone every 10 nucleotides or so, untwists the molecule, and attaches the parental strand to the end of the newly synthesized one. This would synthesize the DNA in short pieces alternating from one strand to the other.

Each of these three models makes a different prediction about the distribution of the “old” parental DNA in molecules formed after replication. To determine which model is actually followed in living cells, the Meselson Stahl experiment made use of the fact that



A summary of the three postulated methods of DNA synthesis

nitrogen (N) is chemically a major constituent of DNA. ^{14}N is by far the most abundant isotope of nitrogen, but DNA with the heavier (but non-radioactive) ^{15}N isotope is also functional.

E. coli were grown for several generations in a medium with ^{15}N . When DNA is extracted from these cells and centrifuged on a salt density gradient, the DNA separates out at the point at which its density equals that of the salt solution. The DNA of the cells grown in ^{15}N medium had a higher density than cells grown in normal ^{14}N medium. After that, *E. coli* cells with only ^{15}N in their DNA were transferred to a ^{14}N medium and were allowed to divide; the progress of cell division was monitored by measuring the optical density of the cell suspension.

DNA was extracted periodically and was compared to pure ^{14}N DNA and ^{15}N DNA. After one replication, the DNA was found to have close to the intermediate density. Since conservative replication would result in equal amounts of DNA of the higher and lower densities (but no DNA of an intermediate density), conservative replication was excluded. However, this result was consistent with both semiconservative and dispersive replication. Semiconservative replication would result in double-stranded DNA with one strand of ^{15}N DNA, and one of ^{14}N DNA, while dispersive replication would result in double-stranded DNA with both strands having mixtures of ^{15}N and ^{14}N DNA, either of which would have appeared as DNA of an intermediate density.

The authors continued to sample cells as replication continued. DNA from cells after two replications had been completed was found to consist of equal amounts of DNA with two different densities, one corresponding to the intermediate density of DNA of cells grown for only one division in ^{14}N medium, the other corresponding to DNA from cells grown exclusively in ^{14}N medium. This was inconsistent with dispersive replication, which would have resulted in a single density, lower than the intermediate density of the one-generation cells, but still higher than cells grown only in ^{14}N DNA medium, as the original ^{15}N DNA would have been split

evenly among all DNA strands. The result was consistent with the semiconservative replication hypothesis.

Replication in prokaryotes

DNA replication in prokaryotes is extensively studied in *E. coli*. It is bi-directional and originates at a single origin of replication.

Primase

In bacteria, primase binds to the DNA helicase forming a complex called the primosome. Primase is activated by DNA helicase where it then synthesizes a short RNA primer, to which new nucleotides can be added by DNA polymerase.

DNA polymerase

In prokaryotic cells there are several kinds of DNA polymerases including:

Pol I: implicated in DNA repair; has 5'→3' polymerase activity, and both 3'→5' exonuclease activity (proofreading) and 5'→3' exonuclease activity (RNA primer removal).

DNA Polymerase I (or Pol I) is an enzyme that participates in the process of DNA replication in prokaryotes. It contains 928 amino acids, and is an example of a processive enzyme – it can sequentially catalyze multiple polymerisations. It was discovered by Arthur Kornberg in 1956, it was the first known DNA polymerase (and, indeed, the first known of any kind of polymerase). It was initially characterized in *E. coli*, although it is ubiquitous in prokaryotes. Pol

I possesses three enzymatic activities: (1) a 5' → 3' (forward) DNA polymerase activity, requiring a 3' primer site and a template strand; (2) a 3' → 5' (reverse) exonuclease activity that mediates proofreading; and (3) a 5' → 3' (forward) exonuclease activity mediating nick translation during DNA repair.

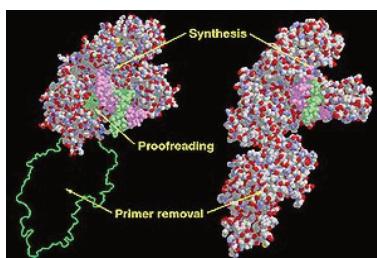
Pol II: involved in repairing damaged DNA; has 3'->5' exonuclease activity. The enzyme is 90 kDa in size and is coded by the polB gene. DNA Pol II can synthesize DNA new base pairs at an average rate of between 40 and 50 nucleotides/second.

Pol III: the main polymerase in bacteria (responsible for elongation); has 3'->5' exonuclease activity (proofreading). Pol III is aided by other machinery at the replication fork. Together this machinery has been called the replisome, composed of two DNA Pol III enzymes that manufacture the leading and lagging strands and sliding DNA clamps that keep the polymerase bound to the DNA.

Uses of DNA polymerase in technology. Klenow fragment

The 5' → 3' exonuclease activity of DNA polymerase I from *E. coli* makes it unsuitable for many molecular biology applications, the “Klenow fragment”, which lacks this activity, can be very useful in research.

The Klenow fragment has been used for tasks such as: (1) Synthesis of double-stranded DNA from single-stranded templates; (2) Filling in (meaning removal of overhangs to create blunt ends) recessed 3' ends of DNA fragments; (3) Digesting away protruding 3' overhangs; (4) Preparation of radioactive DNA probes. The Klenow fragment was also the original enzyme used for greatly amplifying segments of DNA in the



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

polymerase chain reaction (PCR) process, before being replaced by thermostable enzymes such as Taq polymerase.(including

Helicase

The process of semiconservative DNA replication involves the separation of nucleic acid strands at the replication fork. This is accomplished by an enzyme known as helicase which uses the energy from ATP hydrolysis to pull apart the complementary strands. Many other cellular processes besides DNA replication (including transcription, translation, recombination, DNA repair, ribosome biogenesis) also involve the separation of nucleic acid strands and therefore require helicases to separate a DNA double helix or helices that involve RNA. There are many helicases (14 confirmed in *E. coli*, 24 in human cells) that presumably operate in the great variety of processes in which strand separation must be catalyzed. Helicases may process much faster *in vivo* than *in vitro* due to the presence of accessory proteins such as single strand DNA binding proteins that aid in the destabilization of the fork junction and the stabilization of the separated strands.

Topoisomerase

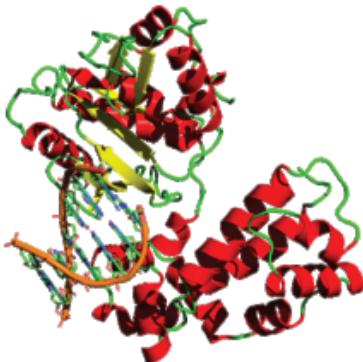
Topoisomerases are enzymes that participate in the overwinding or underwinding of DNA. The winding problem of DNA arises due to the intertwined nature of its double-helical structure. During DNA replication, DNA becomes overwound ahead of a replication fork. If left unabated, this torsion would eventually stop the ability of DNA polymerases to continue down the DNA strand.

In order to prevent and correct these types of topological problems caused by the double helix, topoisomerases bind to DNA

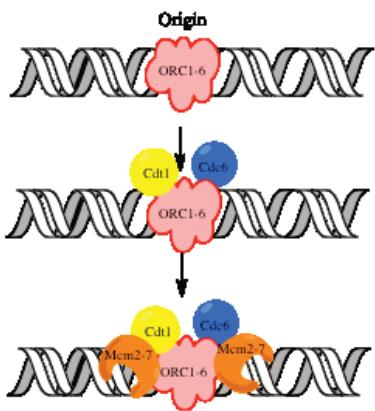
and cut the phosphate backbone of either one or both the DNA strands. This intermediate break allows the DNA to be untangled or unwound, and, at the end of these processes, the DNA backbone is resealed again. Since the overall chemical composition and connectivity of the DNA do not change, the DNA substrate and product are chemical isomers, differing only in their global topology, resulting in the name for these enzymes. Bacterial topoisomerases and human topoisomerases proceed via similar mechanisms for managing DNA supercoils.

Replication in Eukaryotes

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



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



Pre-RC assembly involves the assembly of the ORC subunits, Cdc6 and Cdt1 and the Mcm2-7 complex

DNA replication in eukaryotes occurs only in the S phase of the cell cycle. However, preparation occurs earlier and one of the purposes of cell growth is in preparation of the relatively rapid effects occurring during S-phase. Due to the sheer size of chromosomes in eukaryotes, eukaryotic chromosomes contain multiple origins of replication. An assembly of proteins takes place at each origin. the pre-initiation

replication complex (the pre-RC). The formation of this complex occurs in two stages. Current models hold that it begins with the binding of the origin recognition complex (ORC) to the origin. This complex remains bound to the origin, even after DNA replication occurs. Following the binding of ORC to the origin, a group of proteins that "certify" that the cell is nearing readiness to replicate its DNA (these proteins are known as Cdc6/Cdc18 and Cdt1). Next comes the coordinated loading of the MCM (Mini Chromosome Maintenance) complex to the origin by first binding to ORC and then binding to the MCM complex. The MCM complex is thought to be the major DNA helicase in eukaryotic organisms. Once binding of MCM occurs, a fully ready (fully licensed) pre-RC exists and replication at that particular origin will be forthcoming because at this point a helicase becomes active and unwinding of the parental DNA duplex begins. A protein called Cdc45 recruits all of the DNA replication proteins to the replication fork and DNA synthesis begins.

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

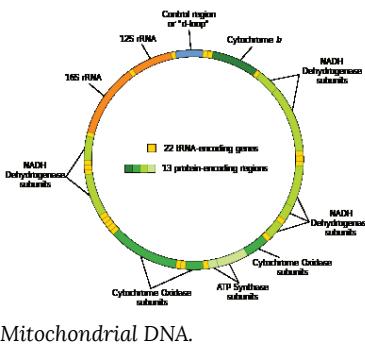
Pol α forms a complex with a small catalytic (PriS) and a large noncatalytic (PriL) subunit, with the Pri subunits acting as a primase (synthesizing an RNA primer), and then with DNA Pol α elongating that primer with DNA nucleotides. After around 20 nucleotides elongation is taken over by Pol ϵ (on the leading strand) and δ (on the lagging strand).

Pol δ : Highly processive and has proofreading 3'->5' exonuclease activity. Thought to be the main polymerase involved in leading strand synthesis, though there is still debate about its role.

Pol ϵ : Also highly processive and has proofreading 3'->5' exonuclease activity. Highly related to pol δ , and thought to be the main polymerase involved in lagging strand synthesis, though there is again still debate about its role.^[22]

Replication in mitochondria

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



DNA repair

DNA damage, due to environmental factors and normal metabolic processes inside the cell, occurs at a rate of 1,000 to 1,000,000 molecular lesions per cell per day. While this constitutes only 0.000165% of the human genome's approximately 6 billion bases (3 billion base pairs), unrepaired lesions in critical genes (such as tumor suppressor genes) can impede a cell's ability to carry out its function and appreciably increase the likelihood of tumor formation.

The vast majority of DNA damage affects the primary structure of the double helix; that is, the bases themselves are chemically modified. These modifications can in turn disrupt the molecules' regular helical structure by introducing non-native chemical bonds or bulky adducts that do not fit in the standard double helix. Unlike proteins and RNA, DNA usually lacks tertiary structure and therefore damage or disturbance does not occur at that level. DNA is, however, supercoiled and wound around "packaging" proteins called histones (in eukaryotes), and both superstructures are vulnerable to the effects of DNA damage.

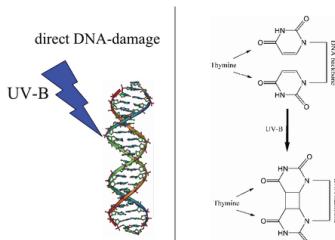
Types of DNA damage

There are five main types of damage to DNA due to endogenous cellular processes: (1) oxidation of bases [e.g. 8-oxo-7,8-dihydroguanine (8-oxoG)] and generation of DNA strand interruptions from reactive oxygen species; (2) alkylation of bases (usually methylation), such as formation of 7-methylguanine, 1-methyladenine, 6-O-Methylguanine; (3) hydrolysis of bases, such as deamination, depurination, and depyrimidination; (4) “bulky adduct formation” (i.e., benzo[a]pyrene diol epoxide-dG adduct); (5) mismatch of bases, due to errors in DNA replication, in which the wrong DNA base is stitched into place in a newly forming DNA strand, or a DNA base is skipped over or mistakenly inserted.

Damage caused by exogenous agents Damage caused by exogenous agents comes in many forms. Some examples are described below. UV-B light causes crosslinking between adjacent cytosine and thymine bases creating **pyrimidine dimers**. This is called direct DNA damage.

UV-A light creates mostly free radicals. The damage caused by free radicals is called indirect DNA damage.

Ionizing radiation such as that created by radioactive decay or in cosmic rays causes breaks in DNA strands. Low-level ionizing radiation may induce irreparable DNA damage (leading to replication and transcriptional errors needed for neoplasia or may trigger viral interactions) leading to pre-mature aging and cancer.



Direct DNA damage: The UV-photon is directly absorbed by the DNA (left). One of the possible reactions from the excited state is the formation of a thymine-thymine cyclobutane dimer (right). The direct DNA damage leads to sunburn, causing an increase in melanin production, thereby leading to a long-lasting tan. However, it is responsible for only 8% of all melanoma.

Thermal disruption at elevated temperature increases the rate of depurination (loss of purine bases from the DNA backbone) and single-strand breaks. For example, hydrolytic depurination is seen in the thermophilic bacteria, which grow in hot springs at 40–80 °C. The rate of depurination (300 purine residues per genome per generation) is too high in these species to be repaired by normal repair machinery, hence a possibility of an adaptive response cannot be ruled out.

Industrial chemicals also play very important role in DNA damage, such as vinyl chloride and hydrogen peroxide, and environmental chemicals such as polycyclic hydrocarbons found in smoke, soot and tar create a huge diversity of DNA adducts- ethenobases, oxidized bases, alkylated phosphotriesters and Crosslinking of DNA just to name a few. **UV damage, alkylation/methylation, X-ray damage and oxidative damage are examples of induced damage.** Spontaneous damage can include the loss of a base, deamination, sugar ring puckering and tautomeric shift.

Sources of damage

DNA damage can be subdivided into **two** main types:

Endogenous damage such as attack by reactive oxygen species produced from normal metabolic byproducts, especially the process of oxidative deamination, and this also includes base mismatches due to replication errors

Exogenous damage caused by external agents such as:

- ultraviolet [UV 200–300 nm] radiation from the sun
- other radiation frequencies, including x-rays and gamma rays
- hydrolysis or thermal disruption
- certain plant toxins
- human-made mutagenic chemicals, especially aromatic compounds that act as DNA intercalating agents

- cancer chemotherapy and radiotherapy
- viruses

Types of mutation

When DNA damages are repaired this can sometimes give rise to a simple one base-pair mutation, described here. (Deletions and translocations can also arise during repair)

Transition In molecular biology, a transition is a point mutation that changes a purine nucleotide to another purine ($A \leftrightarrow G$) or a pyrimidine nucleotide to another pyrimidine ($C \leftrightarrow T$). Approximately two out of three single nucleotide polymorphisms (SNPs) are transitions. Transitions can be caused by oxidative deamination and tautomerization. Although there are twice as many possible transversions, transitions appear more often in genomes, possibly due to the molecular mechanisms that generate them. 5-Methylcytosine is more prone to transition than unmethylated cytosine, due to spontaneous deamination. This mechanism is important because it dictates the rarity of CpG islands.

Transversion In molecular biology, transversion refers to the substitution of a purine for a pyrimidine or vice versa. It can only be reverted by a spontaneous reversion. Because this type of mutation changes the chemical structure dramatically, the consequences of this change tend to be more severe and less common than that of transitions. Transversions can be caused by ionizing radiation and alkylating agents.

DNA repair and disorders

Defects in the NER mechanism are responsible for squally several genetic disorders, including:

Xeroderma pigmentosum: hypersensitivity to sunlight/UV, resulting in increased skin cancer incidence and premature aging

Cockayne syndrome: hypersensitivity to UV and chemical agents

Trichothiodystrophy: sensitive skin, brittle hair and nails Mental retardation often accompanies the latter two disorders, suggesting increased vulnerability of developmental neurons.

Other DNA repair disorders include:

Werner's syndrome: premature aging and retarded growth

Bloom's syndrome: sunlight hypersensitivity, high incidence of malignancies (especially leukemias).

Ataxia telangiectasia: sensitivity to ionizing radiation and some chemical agents

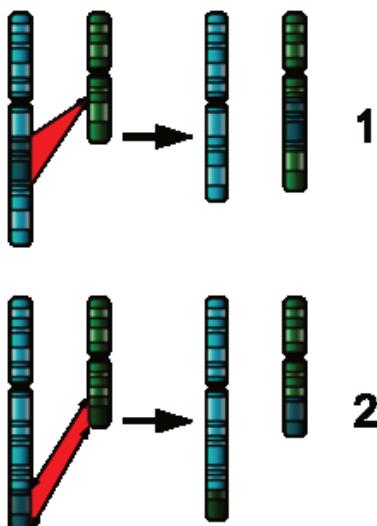
All of the above diseases are often called “**segmental progerias**” (“accelerated aging diseases”) because their victims appear elderly and suffer from aging-related diseases at an abnormally young age, while not manifesting all the symptoms of old age.

Other diseases associated with reduced DNA repair function include **Fanconi's anemia**, hereditary breast cancer and hereditary colon cancer.

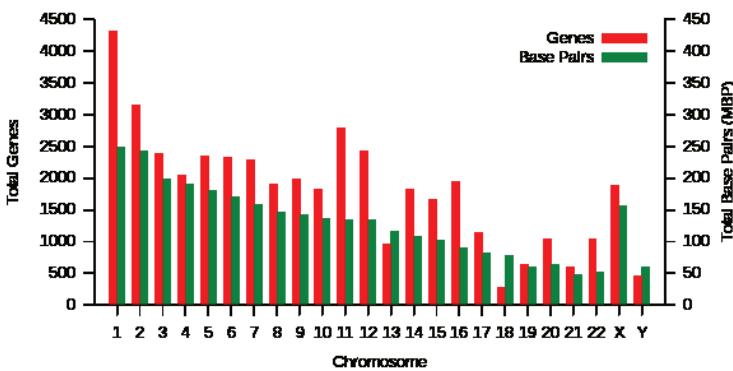
Human Chromosome and Chromosomal aberrations

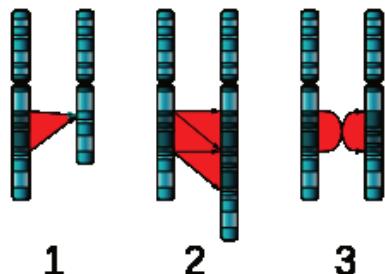
Chromosomes can be divided into two types—autosomes, and sex chromosomes. Certain genetic traits are linked to your sex, and are passed on through the sex chromosomes. The autosomes contain the rest of the genetic hereditary information. All act in the same way during cell division. Human cells have 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.



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

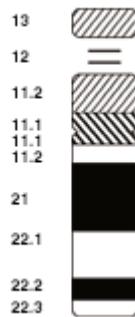




The three major single chromosome mutations; deletion (1), duplication (2) and inversion (3).

Chromosome	Genes	Total bases	Sequenced bases^[28]
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

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:



In Down syndrome, there are three copies of chromosome 21

- **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.

- **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
- **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.

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.

DNA Recombination

Recombination is a process by which a molecule of nucleic acid (usually DNA, but can also be RNA) is broken and then joined to a different one (or in which genetic information is exchanged between two such molecules). Recombination ordinarily occurs between similar molecules of DNA, as in homologous recombination.

Recombination is a common method of DNA repair in both bacteria and eukaryotes. In eukaryotes, recombination also occurs in meiosis, where it facilitates informational exchange and/or chromosomal crossover. The crossover process leads to offspring's having different combinations of genes from those of their parents, and can occasionally produce new chimeric alleles. In organisms with an adaptive immune system, a type of genetic recombination called V(D)J recombination helps immune cells rapidly diversify to recognize and adapt to new pathogens. The shuffling of genes brought about by genetic recombination can have long term advantages, as it is a major engine of genetic variation and also allows sexually reproducing organisms to avoid Muller's ratchet, in which the genomes of an asexual population accumulate deleterious mutations in an irreversible manner. In genetic engineering, recombination can also refer to artificial and deliberate recombination of disparate pieces of DNA, often from different organisms, creating what is called recombinant DNA. A prime



Fig. 61. Scheme to illustrate a method of crossing over of the chromosomes.

Thomas Hunt Morgan's illustration of crossing over (1916)

example of such a use of genetic recombination is gene targeting, which can be used to add, delete or otherwise change an organism's genes. This technique is important to biomedical researchers as it allows them to study the effects of specific genes. Techniques based on genetic recombination are also applied in protein engineering to develop new proteins of biological interest.^[32]

Chromosomal crossover in eukaryotes is an exchange of genetic material between homologous chromosomes. It can occur in one of the final phases of genetic recombination, which occurs during prophase I of meiosis (pachytene). The pairing of homologous chromosomes

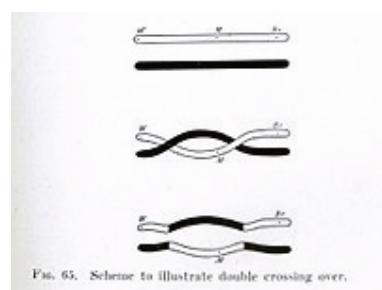
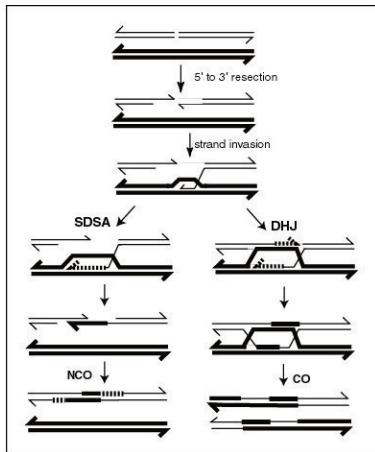


Fig. 65. Scheme to illustrate double crossing over.

during meiosis (synapsis) begins before the synaptonemal complex develops, and is not completed until near the end of prophase I. Crossover usually occurs when matching regions on matching chromosomes break and then reconnect to the other chromosome. Crossing over was described, in theory, by Thomas Hunt Morgan. He relied on the discovery of the Belgian Professor Frans Alfons Janssens of the University of Leuven who described the phenomenon in 1909 and had called it 'chiasmatypie'. The term chiasma is linked if not identical to chromosomal crossover. Morgan immediately saw the great importance of Janssens' cytological interpretation of chiasmata to the experimental results of his research on the heredity of *Drosophila*. The physical basis of crossing over was first demonstrated by Harriet Creighton and Barbara McClintock in 1931.

Meiotic recombination can be initiated by double-stranded breaks that can be introduced into the DNA by the Spo11 protein. In addition, meiotic recombination can be induced in response to spontaneous double strand breaks, possibly caused by reactive oxygen species, carried over from the prior round of synthesis.^[33] One or more exonucleases then digest the 5' ends generated by the double-stranded breaks to produce 3' single-stranded DNA tails (see lowest Figure in this section). The meiosis-specific recombinase Dmc1 and the general recombinase Rad51 coat the single-stranded DNA to form nucleoprotein filaments. The recombinases catalyze invasion of the opposite chromatid by the single-stranded DNA from one end of the break. Next, the 3' end of the invading DNA primes DNA synthesis, causing displacement of the complementary strand.



A current model of meiotic recombination, initiated by a double-strand break or gap, followed by pairing with an homologous chromosome and strand invasion to initiate the recombinational repair process. Repair of the gap can lead to crossover (CO) or non-crossover (NCO) of the flanking regions. CO recombination is thought to occur by the Double Holliday Junction (DHJ) model, illustrated on the right, above. NCO recombinants are thought to occur primarily by the Synthesis Dependent Strand Annealing (SDSA) model, illustrated on the left, above. Most recombination events appear to be the SDSA type.

Crossover recombinants are generated by a process in which the displaced complementary strand subsequently anneals to the single-stranded DNA generated from the other end of the initial double-stranded break (see DHJ pathway on the Figure). The structure that results is a cross-strand exchange, also known as a Holliday junction. The contact between two chromatids that will soon undergo crossing-over is known as a chiasma. The Holliday

junction is a tetrahedral structure which can be ‘pulled’ by other recombinases, moving it along the four-stranded structure (see Double Holliday Junction or DHJ in the Figure).

Gene conversion can result from the repair of a double strand break. Gene conversion involves the unidirectional transfer of genetic sequence information from a ‘donor’ sequence to a highly homologous ‘acceptor’ chromosome. Gene conversion usually occurs by Synthesis Dependent Strand Annealing (SDSA)^{[34][35][36]} illustrated in the lowest Figure in this section. In this model of SDSA DNA repair, a free strand of DNA from the end of a double strand break invades an homologous chromosome, extending itself by replication along the sequence on the complementary strand of DNA of the ‘donor’ chromosome. The extended strand is then retracted from the donor chromosome and pairs with the complementary sequence on the recipient chromosome in a region at the other end of the double strand break (needing about 25 to 50 base pairs of homology).^[34] This allows completion of healing of the double strand break by replication, to complete the duplex structure on the recipient chromosome, from information on the extended strand copied from the donor chromosome. The usual length of a gene conversion tract in mammals is between 200 to 1,000 base pairs.^[37]

During meiosis, gene conversion is most often associated with non-crossover of outside regions (e.g. the SDSA pathway shown in the Figure). Less frequently, gene conversion during meiosis is associated with crossover of outside regions and these events are usually generated by the DHJ pathway. Gene conversion without crossover occurs more frequently than crossover recombination during meiosis in many organisms, often by about a 2 to 1 ratio.^[38] During mitosis, gene conversion is almost the exclusive mode of double strand break repair by homologous recombination.^[36]

Studies of gene conversion have contributed to our understanding of the adaptive function of meiotic recombination. Since gene conversion in most species studied is more frequently of the non-crossover type,^[38] explanations for the adaptive function

of meiotic recombination that focus exclusively on the adaptive benefit of producing new genetic variation seem inadequate to explain the majority of recombination events during meiosis. However, the majority of meiotic recombination events can be explained by the proposal that they are an adaptation for repair of damages in the DNA that is to be passed on to gametes.^{[39][40]}

Genetic recombination is catalyzed by enzymes called recombinases. RecA, the chief recombinase found in *Escherichia coli*, is responsible for the repair of DNA double strand breaks (DSBs). In yeast and other eukaryotic organisms there are two recombinases required for repairing DSBs. The RAD51 protein is employed in both mitotic and meiotic recombination, whereas the DMC1 protein is specific to meiotic recombination.

Nonhomologous recombination Recombinational repair can infrequently occur between DNA sequences that contain no or little sequence homology. This is referred to as nonhomologous recombination.

6.

Technological cycles

Discover something new in molecular biology.

Turn the discovery into a new tool.

Using the new tool, investigate new questions and when successful

Amplification

Viruses

Phage lambda

Other phage

Eukaryotic viruses

Molecular cloning

Vectors and hosts

E. coli and plasmids

Phagemids

Restriction enzymes.

DNA ligase.

DNA sequencing

Sanger sequencing

Marking termination of DNA polymerase.

Next generation sequencing

Human genome project

Applications

Forensics

Personalized medicine

Heritage analysis

PCR amplification

Expression systems

Gene editing

Mutation

Natural mutation

Mutagens

Site-directed mutagenesis. Specifying change in the test tube.

Classical approaches

Microbial selection

Plant breeding

Animal breeding

Recombination

Sexual recombination

Transposons

Viral integration

Recent advances in genome editing

TALENS

CRISPR

References and online resources

The Language of DNA. A free chapter from Jakubowski's Biochemistry Online.

Introduction to Databases. Claire O'Connors Chapter 5 from her free online book Investigations in Molecular Cell Biology.

7.

Case studies

- Ownership of genetic information
- Privacy and personal rights
- Health impacts, good and bad and unknown
- GMOs
- Gene drives on the horizon
- Species eradication
- Molecular Biology Impact Statements

8.

All Creative Commons Licensing, including:

Introduction

The introduction was written by Phil McFadden

Chapter 1

Wikibook “Introduction to Molecular Biology”

New World Encyclopedia Genome

New World Encyclopedia Human genome.

Chapter 2

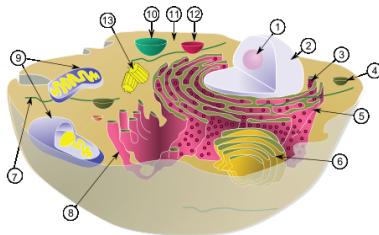
9.

The term Molecular biology was first used by **Warren Weaver** in 1938. Molecular biology is the study of molecular underpinnings of the processes of replication, transcription, translation, and cellular function.

Contents

[hide]

- 1 Macromolecules
 - 1.1 Carbohydrates
 - 1.1.1
 - a.Monosaccharides
 - b.Disaccharide
 - c.Oligosaccharide
 - d.Polysaccharides
 - 1.2 Proteins
 - 1.3 Lipids
 - 1.4 Water
 - 1.5 Noncovalent bond
- 2 pH plays important role in living organism
- 3 Cell The basic unit of life
 - 3.1 Prokaryotes
 - 3.2 Eukaryotic cell
 - 3.3 Plant cell is different from animal cell



A typical animal cell. Within the cytoplasm, the major organelles and cellular structures include: (1) nucleolus (2) nucleus (3) ribosome (4) vesicle (5) rough endoplasmic reticulum (6) Golgi apparatus (7) cytoskeleton (8) smooth endoplasmic reticulum (9) mitochondria (10) vacuole (11) cytosol (12) lysosome (13) centriole.

- 3.4 Origin of Eukaryotic organelles and endosymbiotic theory
 - 3.4.1 Eukaryotic organelles
 - 3.4.2 Prokaryotic organelles
- 4 Macromolecules which are present in the cell membrane
- 5 Facts to be remembered
- 6 Question time
- 7 References

Macromolecules

The term macromolecule was coined by Nobel laureate **Hermann Staudinger** in the 1920s, although his first relevant publication on this field only mentioned high molecular compounds (in excess of 1000 atoms). At that time the phrase polymer as introduced by Berzelius in 1833 had a different meaning from that of today: it simply was another form of isomerism, such as an enzene or acetylene, and had little to do with size. Some examples of organic macromolecules are bio-polymers (carbohydrates, proteins, lipids, nucleic acids) or polymers (plastics, synthetic fiber and rubber).

Carbohydrates

A carbohydrate (ka:bə'haidreɪt/) is an organic compound which has the empirical formula $C_m(H_2O)_n$; that is, consists only of carbon, hydrogen and oxygen, with a hydrogen:oxygen atom ratio of 2:1 (as in water). Carbohydrates can be viewed as hydrates of carbon, hence their name. Structurally however, it is more accurate to view them as polyhydroxy aldehydes and ketones. Historically nutritionists have classified carbohydrates as either simple or complex, however, the exact delineation of these categories is ambiguous. Today, the

term simple carbohydrate typically refers to monosaccharides and disaccharides, and complex carbohydrate means polysaccharides (and oligosaccharides).

a. Monosaccharides

Monosaccharides (from Greek monos: single, sacchar: sugar) are the most basic units of biologically important carbohydrates. They are the simplest form of sugar and are usually colorless, water-soluble, crystalline solids. Some monosaccharides have a sweet taste. Examples of monosaccharides include glucose (dextrose), fructose (levulose), galactose, xylose and ribose. Monosaccharides are the building blocks of disaccharides such as sucrose and polysaccharides (such as cellulose and starch). Further, each carbon atom that supports a hydroxyl group (except for the first and last) is chiral, giving rise to a number of isomeric forms all with the same chemical formula. For instance, galactose and glucose are both aldohexoses, but have different chemical and physical properties.

b. Disaccharide

A disaccharide or biose is the carbohydrate formed when two monosaccharides undergo a condensation reaction which involves the elimination of a small molecule, such as water, from the functional groups only. Like monosaccharides, disaccharides also dissolve in water, taste sweet and are called sugars. The glycosidic bond can be formed between any hydroxyl group on the component monosaccharide. So, even if both component sugars are the same (e.g., glucose), different bond combinations (regiochemistry) and stereochemistry (alpha- or beta-) result in disaccharides that are diastereoisomers with different chemical and physical properties.

Depending on the monosaccharide constituents, disaccharides are sometimes crystalline, sometimes water-soluble, and sometimes sweet-tasting and sticky-feeling.

Disaccharide	Unit 1	Unit 2	Bond
Sucrose (<i>table sugar, cane sugar, beet sugar, or saccharose</i>)	glucose	fructose	$\alpha(1\rightarrow2)$
Lactulose	galactose	fructose	$\beta(1\rightarrow4)$
Lactose (<i>milk sugar</i>)	galactose	glucose	$\beta(1\rightarrow4)$
Maltose	glucose	glucose	$\alpha(1\rightarrow4)$
Trehalose	glucose	glucose	$\alpha(1\rightarrow1)\alpha$
Cellobiose	glucose	glucose	$\beta(1\rightarrow4)$

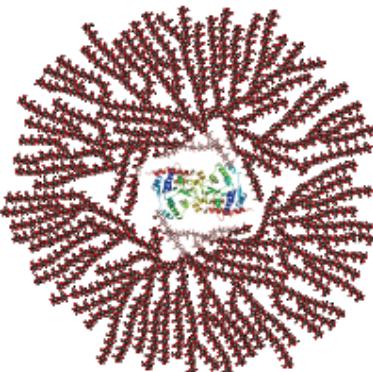
c. Oligosaccharide

An oligosaccharide (from the Greek oligos, a few, and sacchar, sugar) is a saccharide polymer containing typically three to ten component sugars, also known as many as 8 sugars, or polysaccharides. Oligosaccharides can have many functions; for example, they are commonly found on the plasma membrane of animal cells where they can play a role in cell-cell recognition. In general, they are found either O- or N-linked to compatible amino acid side-chains in proteins or to lipid moieties. e.g. **Fructo-oligosaccharides (FOS)**, which are found in many vegetables, consist of short chains of fructose molecules. (Inulin has a much higher degree of polymerization than FOS and is a polysaccharide.) **Galactooligosaccharides (GOS)**, which also occur naturally, consist of short chains of galactose molecules. These compounds can be only partially digested by humans.

d. Polysaccharides

Polysaccharides are polymeric carbohydrate structures, formed of repeating units (either mono- or disaccharides) joined together by glycosidic bonds. These structures are often linear, but may contain various degrees of branching. Polysaccharides are often quite heterogeneous, containing slight modifications of the repeating unit.

Depending on the structure, these macromolecules can have distinct properties from their monosaccharide building blocks. They may be amorphous or even insoluble in water. Starches are glucose polymers in which glucopyranose units are bonded by alpha-linkages. It is made up of a mixture of Amylose (15–20%) and Amylopectin (80–85%). Amylose consists of a linear chain of several hundred glucose molecules and Amylopectin is a branched molecule made of several thousand glucose units (every chain 24–30 glucose unit). Starches are insoluble in water. They can be digested by hydrolysis, catalyzed by enzymes called amylases, which can break the alpha-linkages (glycosidic bonds). Humans and other animals have amylases, so they can digest starches. Potato, rice, wheat, and maize are major sources of starch in the human diet. The formation of starches are the way that plants store glucose. Glycogen is a polysaccharide that is found in animals and is composed of a branched chain of glucose residues. It is stored in liver and muscles. Chitin is one of many naturally occurring polymers. It is one of the most abundant natural materials in the world. Over time it is bio-degradable in the natural environment. Its breakdown may be catalyzed by enzymes called chitinases, secreted by microorganisms such as bacteria and fungi, and produced by



Glycogen.

some plants. **Arabinoxylans** are the copolymers of two pentose sugars – arabinose and xylose.

Proteins

Proteins are polymer of amino acids linked together by peptide bonds. Amino acids can be divided into two group **essential amino acids** and **non-essential amino acids**. Proteins and carbohydrates contain 4 kcal/gram as opposed to lipids which contain 9 kcal/gram. The liver, and to a much lesser extent the kidneys, can convert amino acids used by cells in protein biosynthesis into glucose by a process known as gluconeogenesis. The essential amino acids, which must be obtained from external sources (food), are leucine, isoleucine, valine, lysine, threonine, tryptophan, methionine, phenylalanine and histidine. On the other hand, non-essential amino acids are synthesized in our body from other amino acids. The non-essential amino acids are arginine, alanine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, proline, serine, and tyrosine. Proteins (/'proʊti:nz/; also known as polypeptides) are organic compounds made of amino acids arranged in a linear chain and folded into a globular or fibrous form. The amino acids in a polymer are joined together by the peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. Proteins were first described by the Dutch chemist Gerhardus Johannes Mulder and named by the Swedish chemist Jöns Jakob Berzelius in 1838.

Lipids

Lipids are a broad group of naturally occurring molecules which includes fats, waxes, sterols, fat-soluble vitamins (such as vitamins A, D, E and K), monoglycerides, diglycerides, phospholipids, and others. The main biological functions of lipids include energy storage, as structural components of cell membranes, and as important signaling molecules. Cells contain about 70% of water.

Lipids in membrane Eukaryotic cells are compartmentalized into membrane-bound organelles which carry out different biological functions. The glycerophospholipids are the main structural component of biological membranes, such as the cellular plasma membrane and the intracellular membranes of organelles; in animal cells the plasma membrane physically separates the intracellular components from the extracellular environment. The glycerophospholipids are amphipathic molecules (containing both hydrophobic and hydrophilic regions) that contain a glycerol core linked to three fatty acid-derived "tails" by ester linkages and to one "head" group by a phosphate ester linkage. While glycerophospholipids are the major component of biological membranes, other non-glyceride lipid components such as sphingomyelin and sterols (mainly cholesterol in animal cell membranes) are also found in biological membranes. In plants and algae, the galactosyldiacylglycerols, and sulfoquinovosyldiacylglycerol, which lack a phosphate group, are important components of membranes of chloroplasts and related organelles and are the most abundant lipids in photosynthetic tissues, including those of higher plants, algae and certain bacteria. Bilayers have been found to exhibit high levels of birefringence which can be used to probe the degree of order (or disruption) within the bilayer using techniques such as dual polarisation interferometry.

A biological membrane is a form of lipid bilayer. The formation

of lipid bilayers is an energetically preferred process when the glycerophospholipids described above are in an aqueous environment. In an aqueous system, the polar heads of lipids align towards the polar, aqueous environment, while the hydrophobic tails minimize their contact with water and tend to cluster together, forming a vesicle; depending on the concentration of the lipid, this biophysical interaction may result in the formation of micelles, liposomes, or lipid bilayers. Other aggregations are also observed and form part of the polymorphism of amphiphile (lipid) behavior. Phase behavior is an area of study within biophysics and is the subject of current academic research. Micelles and bilayers form in the polar medium by a process known as the hydrophobic effect. When dissolving a lipophilic or amphiphilic substance in a polar environment, the polar molecules (i.e., water in an aqueous solution) become more ordered around the dissolved lipophilic substance, since the polar molecules cannot form hydrogen bonds to the lipophilic areas of the amphiphile. So in an aqueous environment, the water molecules form an ordered “clathrate” cage around the dissolved lipophilic molecule.

Role of lipid in signalling

In recent years, evidence has emerged showing that lipid signaling is a vital part of the cell signaling. Lipid signaling may occur via activation of G protein-coupled or nuclear receptors, and members of several different lipid categories have been identified as signaling molecules and cellular messengers. These include sphingosine-1-phosphate, a sphingolipid derived from ceramide that is a potent messenger molecule involved in regulating calcium mobilization, cell growth, and apoptosis; diacylglycerol (DAG) and the phosphatidylinositol phosphates (PIPs), involved in calcium-mediated activation of protein kinase C; the prostaglandins, which are one type of fatty-acid derived eicosanoid involved in inflammation and immunity; the steroid hormones such as estrogen, testosterone and cortisol, which modulate a host of functions such as reproduction, metabolism and blood pressure;

and the oxysterols such as 25-hydroxy-cholesterol that are liver X receptor agonists.

Water

We know water is essential to all life as we know it so much so that every cell is nearly 70% of water. We also know that one molecule of water is made up of two hydrogen (H) atoms which are covalently bonded to a single oxygen(O) atom (water's chemical formula H₂O).

Noncovalent bond

A noncovalent bond is a type of chemical bond, typically between macromolecules, that does not involve the sharing of pairs of electrons, but rather involves more dispersed variations of electromagnetic interactions. The noncovalent bond is the dominant type of bond between supermolecules in supramolecular chemistry. Noncovalent bonds are critical in maintaining the three-dimensional structure of large molecules, such as proteins and nucleic acids, and are involved in many biological processes in which large molecules bind specifically but transiently to one another. The energy released in the formation of noncovalent bonds is on the order of **1-5 kcal per mol**. There are four commonly mentioned types of non-covalent interactions: **hydrogen bonds, ionic bonds, van der Waals forces, and hydrophobic interactions**. The noncovalent interactions hold together the two strands of DNA in the double helix, stabilize secondary and tertiary structures of proteins, and enable enzyme-substrate binding and antibody-antigen association.

Intramolecular noncovalent interactions are largely responsible for the secondary and tertiary structure of proteins and therefore the

protein's function in the mechanisms of life. Intermolecular noncovalent interactions are responsible for protein complexes (quaternary structure) where two or more proteins function in a coherent mechanism.

Most drugs work by noncovalently interacting with biomolecules such as proteins or RNA. Relatively few drugs actually form covalent bonds with the biomolecules they interact with; instead, they interfere with or activate some biological mechanism through noncovalently interacting in very specific locations on specific biomolecules which present the perfect combination of noncovalent binding partners in just the right geometry.

Hydrogen bonding

The best example of a hydrogen bond is found between water molecules. Water molecules contain two hydrogen atoms and one oxygen atom. Two molecules of water can form a hydrogen bond between them. Inside the cell hydrogen bonding also plays an important role in determining the 3D structures of proteins and nucleic bases. In these macromolecules, bonding between parts of the same macromolecule cause it to fold into a specific shape, which helps determine the molecule's physiological or biochemical function. The double helical structure of DNA, for example, is due to hydrogen bonding between the base pairs, which link one complementary strand to the other and enable replication. Hydrogen bonds are also important in the structure of macromolecule cellulose. We should also remember that the hydrogen bond is stronger than a van der Waals interaction, but generally weaker than ionic bonds.

pH plays important role in living organism

As we know very well in chemistry, pH is a measure of the acidity or basicity of an aqueous solution. Pure water is said to be neutral, with a pH close to 7.0 at 25 °C (77 °F). Solutions with a pH less than 7

are said to be acidic and solutions with a pH greater than 7 are basic or alkaline. pH measurements are important in medicine, biology, chemistry, food science, environmental science, oceanography, civil engineering and many other applications.

In a solution pH approximates but is not equal to $p[H]$, the negative logarithm (base 10) of the molar concentration of dissolved hydronium ions (H_3O^+); a low pH indicates a high concentration of hydronium ions, while a high pH indicates a low concentration. Crudely, this negative of the logarithm matches the number of places behind the decimal point, so for example 0.1 molar hydrochloric acid should be near pH 1 and 0.0001 molar HCl should be near pH 4 (the base 10 logarithms of 0.1 and 0.0001 being -1, and -4, respectively). Pure (de-ionised) water is neutral, and can be considered either a very weak acid or a very weak base (center of the 0 to 14 pH scale), giving it a pH of 7 (at 25 °C (77 °F)), or 0.0000001 M H^+ . For an aqueous solution to have a higher pH, a base must be dissolved in it, which binds away many of these rare hydrogen ions. Hydrogen ions in water can be written simply as H^+ or as hydronium (H_3O^+) or higher species (e.g. $H_9O_4^+$) to account for solvation, but all describe the same entity.

Most of the Earth's freshwater surface bodies are slightly acidic due to the abundance and absorption of carbon dioxide; in fact, for millennia in the past most fresh water bodies have long existed at a slightly acidic pH level. However, pH is not precisely $p[H]$, but takes into account an activity factor. This represents the tendency of hydrogen ions to interact with other components of the solution, which affects among other things the electrical potential read using a pH meter. As a result, pH can be affected by the ionic strength of a solution – for example, the pH of a 0.05 M potassium hydrogen phthalate solution can vary by as much as 0.5 pH units as a function of added potassium chloride, even though the added salt is neither acidic nor basic. pH also play important role in living organism.

The pH of different cellular compartments, body fluids, and organs is usually tightly regulated in a process called acid-base homeostasis. The pH of blood is usually slightly basic with a value

of pH 7.365. This value is often referred to as physiological pH in biology and medicine. Plaque can create a local acidic environment that can result in tooth decay by demineralisation. Enzymes and other proteins have an optimum pH range and can become inactivated or denatured outside this range. The most common disorder in acid-base homeostasis is acidosis, a condition in which there is an acid overload in the body, generally defined by pH falling below 7.35. In the blood, pH can be estimated from known base excess (be) and bicarbonate concentration (HCO_3) by the following equation:

$$\text{pH} = \frac{\text{be} - 0.963\text{HCO}_3 + 124}{13.77}$$

Cell The basic unit of life

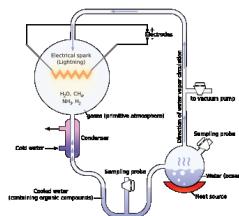
The cell is the functional basic unit of life. It was discovered by **Robert Hooke** and is the functional unit of all known living organisms. It is the smallest unit of life that is classified as a living thing, and is often called the building block of life. Some organisms, such as most bacteria, are unicellular (consist of a single cell). Other organisms, such as humans and birds, are multicellular. Humans have about 100 trillion or 10^{14} cells; a typical cell size is 10 μm and a typical cell mass is 1 nanogram. **The largest cells are about 135 μm in the anterior horn in the spinal cord while granule cells in the cerebellum, the smallest, can be some 4 μm** and the longest cell can reach from the toe to the lower brain stem (Pseudounipolar cells).

The largest known cells are unfertilised ostrich egg cells which weigh 3.3 pounds. In 1835, before the final cell theory was developed, Jan Evangelista Purkyně observed small “granules” while looking at the plant tissue through a microscope. The cell theory, first developed in 1839 by Matthias Jakob Schleiden and Theodor Schwann, states that all organisms are composed of one or more cells, that all cells come from preexisting cells, that vital functions of an organism occur within cells, and that all cells contain the

hereditary information necessary for regulating cell functions and for transmitting information to the next generation of cells. **The word cell comes from the Latin cellula, meaning, a small room.** The descriptive term for the smallest living biological structure was coined by Robert Hooke in a book he published in 1665 when he compared the cork cells he saw through his microscope to the small rooms monks lived in. There are two types of cells: **eukaryotic and prokaryotic**. Prokaryotic cells are usually independent, while eukaryotic cells are often found in multicellular organisms.

Origin of life and Miller's experiment

Earth's early atmosphere Some evidence suggests that Earth's original atmosphere might have contained fewer of the reducing molecules than was thought at the time of the Miller–Urey experiment. There is abundant evidence of major volcanic eruptions 4 billion years ago, which would have released carbon dioxide, nitrogen, hydrogen sulfide (H_2S), and sulfur dioxide (SO_2) into the atmosphere. Experiments using these gases in addition to the ones in



The experiment

the original Miller–Urey experiment have produced more diverse molecules. The experiment created a mixture that was racemic (containing both L and D enantiomers) and experiments since have shown that “in the lab the two versions are equally likely to appear.”^[1] However, in nature, L amino acids dominate; later experiments have confirmed disproportionate amounts of L or D oriented enantiomers are possible.^[2]

Originally it was thought that the primitive secondary atmosphere contained mostly ammonia and methane. However, it is likely that most of the atmospheric carbon was CO₂ with perhaps some CO and the nitrogen mostly N₂. In practice gas mixtures containing CO, CO₂, N₂, etc. give much the same products as those containing CH₄ and NH₃ so long as there is no O₂. The hydrogen atoms come mostly from water vapor. In fact, in order to generate aromatic amino acids under primitive earth conditions it is necessary to use less hydrogen-rich gaseous mixtures. Most of the natural amino acids, hydroxyacids, purines, pyrimidines, and sugars have been made in variants of the Miller experiment.^[3]

More recent results may question these conclusions. The University of Waterloo and University of Colorado conducted simulations in 2005 that indicated that the early atmosphere of Earth could have contained up to 40 percent hydrogen—implying a possibly much more hospitable environment for the formation of prebiotic organic molecules. The escape of hydrogen from Earth's atmosphere into space may have occurred at only one percent of the rate previously believed based on revised estimates of the upper atmosphere's temperature.^[4] One of the authors, Owen Toon notes: "In this new scenario, organics can be produced efficiently in the early atmosphere, leading us back to the organic-rich soup-in-the-ocean concept... I think this study makes the experiments by Miller and others relevant again." Outgassing calculations using a chondritic model for the early earth complement the Waterloo/Colorado results in re-establishing the importance of the Miller–Urey experiment.^[5]

Conditions similar to those of the Miller–Urey experiments are present in other regions of the solar system, often substituting ultraviolet light for lightning as

the energy source for chemical reactions. The Murchison meteorite that fell near Murchison, Victoria, Australia in 1969 was found to contain over 90 different amino acids, nineteen of which are found in Earth life. Comets and other icy outer-solar-system bodies are thought to contain large amounts of complex carbon compounds (such as tholins) formed by these processes, darkening surfaces of these bodies.^[6] The early Earth was bombarded heavily by comets, possibly providing a large supply of complex organic molecules along with the water and other volatiles they contributed. This has been used to infer an origin of life outside of Earth: the panspermia hypothesis. The **Miller and Urey experiment**^[7] (or **Urey–Miller experiment**)^[8] was an experiment that simulated hypothetical conditions thought at the time to be present on the early Earth, and tested for the occurrence of chemical origins of life. Specifically, the experiment tested Alexander Oparin's and J. B. S. Haldane's hypothesis that conditions on the primitive Earth favored chemical reactions that synthesized [organic] compounds from inorganic precursors. Considered to be the classic experiment on

the origin of life, it was conducted in 1952^[9] and published in 1953 by Stanley Miller and Harold Urey at the University of Chicago.^{[10][11][12]}

After Miller's death in 2007, scientists examining sealed vials preserved from the original experiments were able to show that there were actually well over 20 different amino acids produced in Miller's original experiments. That is considerably more than what Miller originally reported, and more than the 20 that naturally occur in life. Moreover, some evidence suggests that Earth's original atmosphere might have had a different composition than the gas used in the Miller–Urey experiment. There is abundant evidence of major volcanic eruptions 4 billion years ago, which would have released carbon dioxide, nitrogen, hydrogen sulfide (H_2S), and sulfur dioxide (SO_2) into the atmosphere. Experiments using these gases in addition to the ones in the original Miller–Urey experiment have produced more diverse molecules.^[1]

Experiment

The experiment used water (H_2O), methane (CH_4), ammonia (NH_3), and

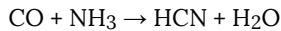
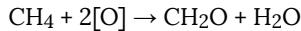
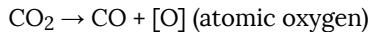
hydrogen (H_2). The chemicals were all sealed inside a sterile array of glass tubes and flasks connected in a loop, with one flask half-full of liquid water and another flask containing a pair of electrodes. The liquid water was heated to induce evaporation, sparks were fired between the electrodes to simulate lightning through the atmosphere and water vapor, and then the atmosphere was cooled again so that the water could condense and trickle back into the first flask in a continuous cycle.

At the end of one week of continuous operation, Miller and Urey observed that as much as 10–15% of the carbon within the system was now in the form of organic compounds. Two percent of the carbon had formed amino acids that are used to make proteins in living cells, with glycine as the most abundant. Sugars, liquids, were also formed. Nucleic acids were not formed within the reaction. But the common 20 amino acids were formed, but in various concentrations.

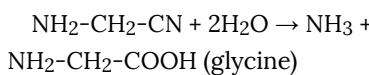
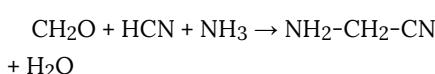
In an interview, Stanley Miller stated: “Just turning on the spark in a basic prebiotic experiment will yield 11 out of 20 amino acids.”^[13]

As observed in all subsequent experiments, both left-handed (L) and right-handed (D) optical isomers were created in a racemic mixture. The original experiment remains today under the care of Miller and Urey's former student Professor Jeffrey Bada at the University of California, San Diego, Scripps Institution of Oceanography.^[14]

One-step reactions among the mixture components can produce hydrogen cyanide (HCN), formaldehyde (CH_2O),^[15] and other active intermediate compounds (acetylene, cyanoacetylene, etc.):



The formaldehyde, ammonia, and HCN then react by Strecker synthesis to form amino acids and other biomolecules:



Furthermore, water and formaldehyde

can react via Butlerov's reaction to produce various sugars like ribose.

Other experiments This experiment inspired many others. In 1961, Joan Oró found that the nucleotide base adenine could be made from hydrogen cyanide (HCN) and ammonia in a water solution. His experiment produced a large amount of adenine, which molecules were formed from 5 molecules of HCN.^[16] Also, many amino acids are formed from HCN and ammonia under these conditions.^[17] Experiments conducted later showed that the other RNA and DNA nucleobases could be obtained through simulated prebiotic chemistry with a reducing atmosphere.^[18]

There also had been similar electric discharge experiments related to the origin of life contemporaneous with Miller–Urey. An article in *The New York Times* (March 8, 1953:E9), titled “Looking Back Two Billion Years” describes the work of Wollman (William) M. MacNevin at The Ohio State University, before the Miller Science paper was published in May 1953. MacNevin was passing 100,000 volt sparks through methane and water vapor and produced “resinous solids” that were “too complex

for analysis." The article describes other early earth experiments being done by MacNevin. It is not clear if he ever published any of these results in the primary scientific literature.^[citation needed]

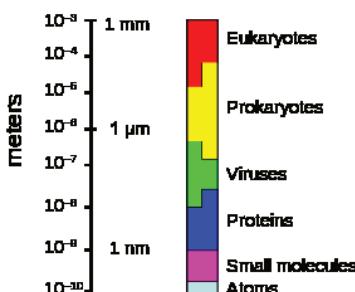
K. A. Wilde submitted a paper to *Science* on December 15, 1952, before Miller submitted his paper to the same journal on February 14, 1953. Wilde's paper was published on July 10, 1953.^[19] Wilde used voltages up to only 600 V on a binary mixture of carbon dioxide (CO₂) and water in a flow system. He observed only small amounts of carbon dioxide reduction to carbon monoxide, and no other significant reduction products or newly formed carbon compounds. Other researchers were studying UV-photolysis of water vapor with carbon monoxide. They have found that various alcohols, aldehydes and organic acids were synthesized in reaction mixture.^[20]

More recent experiments by chemist Jeffrey Bada at Scripps Institution of Oceanography (in La Jolla, CA) were similar to those performed by Miller. However, Bada noted that in current models of early Earth conditions, carbon dioxide and nitrogen (N₂) create nitrite]s, which

destroy amino acids as fast as they form. However, the early Earth may have had significant amounts of iron and carbonate minerals able to neutralize the effects of the nitrites. When Bada performed the Miller-type experiment with the addition of iron and carbonate minerals, the products were rich in amino acids. This suggests the origin of significant amounts of amino acids may have occurred on Earth even with an atmosphere containing carbon dioxide and nitrogen.^[21]

Prokaryotes

Prokaryotes are single-cell organisms that do not have a nucleus, mitochondria, or any other membrane-bound organelles. In other words, neither their DNA nor any of their other sites of metabolic activity are collected together in a discrete membrane-enclosed area. Instead, everything is openly accessible within the cell, some of which is free-floating. A distinction between



The sizes of prokaryotes relative to other organisms and biomolecules

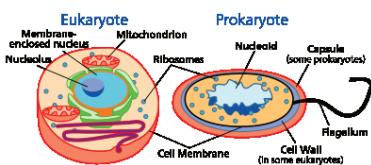
prokaryotes and eukaryotes (meaning true kernel, also spelled “eucaryotes”) is that eukaryotes do have “true” nuclei containing their DNA. Unlike prokaryotes, eukaryotic organisms may be unicellular, as in amoebae, or multicellular, as in plants and animals. The difference between the structure of prokaryotes and eukaryotes is so great that it is sometimes considered to be the most important distinction among groups of organisms. The cell structure of prokaryotes differs greatly from that of eukaryotes. The defining characteristic is the absence of a nucleus. Also the size of Ribosomes in prokaryotes is smaller than that in eukaryotes, which is now where respiration takes place. The genomes of prokaryotes are held within an irregular DNA/protein complex in the cytosol called the nucleoid, which lacks a nuclear envelope.

In general, prokaryotes lack the following membrane-bound cell compartments: mitochondria and chloroplasts. Instead, processes such as oxidative phosphorylation and photosynthesis take place across the prokaryotic plasma membrane. However, prokaryotes do possess some internal structures, such as cytoskeletons, and the bacterial order Planctomycetes have a membrane around their nucleoid and contain other membrane-bound cellular structures. Both eukaryotes and prokaryotes contain large RNA/protein structures called ribosomes, which produce protein. Prokaryotes are usually much smaller than eukaryotic cells. Prokaryotes also differ from eukaryotes in that they contain only a single loop of stable chromosomal DNA stored in an area named the nucleoid, whereas eukaryote DNA is found on tightly bound and organized chromosomes. Although some eukaryotes have satellite DNA structures called plasmids, in general these are regarded as a prokaryote feature, and many important genes in prokaryotes are stored on plasmids. Prokaryotes have a larger surface-area-to-volume ratio giving them a higher metabolic rate, a higher growth rate, and, as a consequence, a shorter generation time compared to Eukaryotes. A criticism of this classification is that the word “prokaryote” is based on what these organisms are not (they are not eukaryotic), rather than what they are (either archaea or bacteria).

In 1977, Carl Woese proposed dividing prokaryotes into the Bacteria and Archaea (originally Eubacteria and Archaeabacteria) because of the major differences in the structure and genetics between the two groups of organisms. This arrangement of Eukaryota (also called “Eukarya”), Bacteria, and Archaea is called the three-domain system, replacing the traditional two-empire system.

Eukaryotic cell

The cells of eukaryotes (left) and prokaryotes (right). The origin of the eukaryotic cell was a milestone in the evolution of life, since they include all complex cells and almost all multi-cellular organisms. The



The cells of eukaryotes (left) and prokaryotes (right)

timing of this series of events is hard to determine; Knoll (2006) suggests they developed approximately 1.6 – 2.1 billion years ago. Some acritarchs are known from at least 1,650 million years ago, and the possible alga Grypania has been found as far back as 2,100 million years ago. Fossils that are clearly related to modern groups start appearing around 1.2 billion years ago, in the form of a red alga, though recent work suggests the existence of fossilized filamentous algae in the Vindhya basin dating back to 1.6 to 1.7 billion years ago. Biomarkers suggest that at least stem eukaryotes arose even earlier. The presence of steranes in Australian shales indicates that eukaryotes were present 2.7 billion years ago.

There are many different types of eukaryotic cells, though animals and plants are the most familiar eukaryotes, and thus provide an excellent starting point for understanding eukaryotic structure. Fungi and many protists have some substantial differences, however.

Animal cell

An animal cell is a form of eukaryotic cell that makes up many tissues in animals. The animal cell is distinct from other eukaryotes, most notably plant cells, as they lack cell walls and chloroplasts, and they have smaller vacuoles. Due to the lack of a rigid cell wall, animal cells can adopt a variety of shapes, and a phagocytic cell can even engulf other structures.

There are many different cell types. For instance, there are approximately 210 distinct cell types in the adult human body.

Plant cell

Plant cells are quite different from the cells of the other eukaryotic organisms. Their distinctive features are: A large central vacuole (enclosed by a membrane, the tonoplast), which maintains the cell's turgor and controls movement of molecules between the cytosol and sap. A primary cell wall containing cellulose, hemicellulose and pectin, deposited by the protoplast on the outside of the cell membrane; this contrasts with the cell walls of fungi, which contain chitin, and the cell envelopes of prokaryotes, in which peptidoglycans are the main structural molecules. The plasmodesmata, linking pores in the cell wall that allow each plant cell to communicate with other adjacent cells; this is different from the functionally analogous system of gap junctions between animal cells. Plastids, especially chloroplasts that contain chlorophyll, the pigment that gives plants their green color and allows them to perform photosynthesis. Higher plants, including conifers and flowering plants (Angiospermae) lack the flagellae and centrioles that are present in animal cells.

Fungal cell

Fungal cells are most similar to animal cells, with the following exceptions: A cell wall that contains chitin. Less definition between cells; the hyphae of higher fungi have porous partitions called septa, which allow the passage of cytoplasm, organelles, and, sometimes, nuclei. Primitive fungi have few or no septa, so each organism is essentially a giant multinucleate supercell; these fungi are described as coenocytic. Only the most primitive fungi, chytrids, have flagella.

Other eukaryotic cells

Eukaryotes are a very diverse group, and their cell structures are equally diverse. Many have cell walls; many do not. Many have chloroplasts, derived from primary, secondary, or even tertiary endosymbiosis; and many do not. Some groups have unique structures, such as the cyanelles of the glaucophytes, the haptonema of the haptophytes, or the ejectisomes of the cryptomonads. Other structures, such as pseudopods, are found in various eukaryote groups in different forms, such as the lobose amoebozoans or the reticulose foraminiferans.

Table 1: Comparison of features of Prokaryotic and Eukaryotic cells

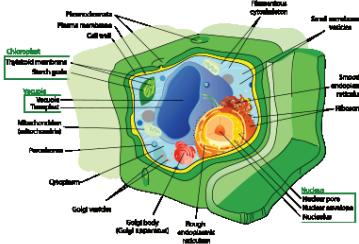
	Prokaryotes	Eukaryotes
Typical organisms	Bacteria, archaea	Protists, Fungi, Plants, Animals
Typical size	~ 1–10 µm	~ 10–100 µm (sperm cells, apart from the tail, are smaller)
Type of nucleus	nucleoid region; no real nucleus	real nucleus surrounded by double membrane
DNA	circular (usually)	linear molecules (chromosomes) with histone proteins
RNA-/protein-synthesis	coupled in cytoplasm	RNA-synthesis inside the nucleus protein synthesis in cytoplasm
Ribosomes	50S+30S	60S+40S
Cytoplasmatic structure	very few structures	highly structured by endomembranes and a cytoskeleton
Cell movement	flagella made of flagellin	flagella and cilia containing microtubules; lamellipodia and filopodia containing actin
Mitochondria	none	one to several thousand (though some lack mitochondria)
Chloroplasts	none	in algae and plants
Organization	usually single cells	single cells, colonies, higher multicellular organisms with specialized cells
Cell division	Binary fission (simple division)	Mitosis (fission or budding) Meiosis

Plant cell is different from animal cell

Plant cells are eukaryotic cells that differ in several key respects from the cells of other eukaryotic organisms. Their distinctive features include: A large central **vacuole**, a water-filled volume enclosed by a membrane known as the **tonoplast** maintains the cell's

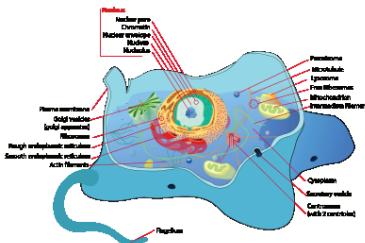
turgor, controls movement of molecules between the cytosol and sap, stores useful material and digests waste proteins and organelles.

A **cell wall** composed of cellulose and hemicellulose, pectin and in many cases lignin, are secreted by the **protoplast** on the outside of the cell membrane. This contrasts with the cell walls of fungi (which are made of chitin), and of bacteria, which are made of peptidoglycan. Specialised cell-cell communication pathways known as plasmodesmata, pores in the primary cell wall through which the plasmalemma and endoplasmic reticulum of adjacent cells are continuous.



Structure of a typical plant cell

pigments. As in mitochondria, which have a genome encoding 37 genes, plastids have their own genomes of about 100-120 unique



Structure of a typical animal cell

Plastids, the notables one being the **chloroplasts**, which contain chlorophyll and the biochemical systems for light harvesting and photosynthesis, but also amyloplasts specialized for starch storage, elaioplasts specialized for fat storage, and chromoplasts specialized for synthesis and storage of

genes and, it is presumed, arose as prokaryotic endosymbionts living in the cells of an early eukaryotic ancestor of the land plants and algae.

Unlike animal cells, plant cells are stationary. Cell division by construction of a phragmoplast as a template for building a cell plate late in cytokinesis is characteristic of land plants and a few groups of algae, the notable one being the Charophytes and the Order Trentepohliales. The sperm of bryophytes have flagellae similar to those in animals, but higher plants, (including Gymnosperms and flowering plants) **lack the flagellae and centrioles that are present in animal cells.**

Table 2: Comparison of structures between animal and plant cells

	Typical animal cell	Typical plant cell
Organelles	<ul style="list-style-type: none">• Nucleus<ul style="list-style-type: none">◦ Nucleolus (within nucleus)• Rough endoplasmic reticulum (ER)• Smooth ER• Ribosomes• Cytoskeleton• Golgi apparatus• Cytoplasm• Mitochondria• Vesicles• Lysosomes• Centrosome<ul style="list-style-type: none">◦ Centrioles	<ul style="list-style-type: none">• Nucleus<ul style="list-style-type: none">◦ Nucleolus (within nucleus)• Rough ER• Smooth ER• Ribosomes• Cytoskeleton• Golgi apparatus (dictiosomes)• Cytoplasm• Mitochondria• Plastids and its derivatives• Vacuole(s)• Cell wall

Origin of Eukaryotic organelles and

endosymbiotic theory

The endosymbiotic (from the Greek: *endo-* meaning inside and *-symbiosis* meaning cohabiting) theory was first articulated by the **Russian botanist Konstantin Mereschkowski in 1905**. Mereschkowski was familiar with work by botanist Andreas Schimper, who had observed in 1883 that the division of chloroplasts in green plants closely resembled that of free-living cyanobacteria, and who had himself tentatively proposed (in a footnote) that green plants had arisen from a symbiotic union of two organisms. Ivan Wallin extended the idea of an endosymbiotic origin to mitochondria in the 1920s. These theories were initially dismissed or ignored. More detailed electron microscopic comparisons between cyanobacteria and chloroplasts (for example studies by Hans Ris), combined with the discovery that plastids and mitochondria contain their own DNA (which by that stage was recognized to be the hereditary material of organisms) led to a resurrection of the idea in the 1960s. The endosymbiotic theory was advanced and substantiated with microbiological evidence by Lynn Margulis in a 1967 paper, *The Origin of Mitosing Eukaryotic Cells*.

In her 1981 work *Symbiosis in Cell Evolution* she argued that eukaryotic cells originated as communities of interacting entities, including endosymbiotic spirochaetes that developed into eukaryotic flagella and cilia. This last idea has not received much acceptance, because flagella lack DNA and do not show ultrastructural similarities to bacteria or archaea. According to Margulis and Dorion Sagan, “Life did not take over the globe by combat, but by networking” (i.e., by cooperation). The possibility that peroxisomes may have an endosymbiotic origin has also been considered, although they lack DNA. Christian de Duve proposed that they may have been the first endosymbionts, allowing cells to withstand growing amounts of free molecular oxygen in the Earth’s atmosphere. However, it now appears that they may be formed *de novo*, contradicting the idea that they have a symbiotic origin.

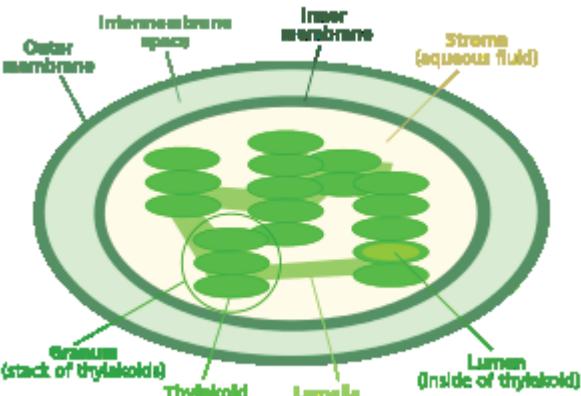
It is believed that over millennia these endosymbionts transferred some of their own DNA to the host cell's nucleus during the evolutionary transition from a symbiotic community to an instituted eukaryotic cell (called "serial endosymbiosis"). This hypothesis is thought to be possible because it is known today from scientific observation that transfer of DNA occurs between bacteria species, even if they are not closely related. Bacteria can take up DNA from their surroundings and have a limited ability to incorporate it into their own genome.

Eukaryotic organelles

Eukaryotes are one of the structurally complex cell type, and by definition are in part organized by smaller interior compartments, that are themselves enclosed by lipid membranes that resemble the outermost cell membrane. The larger organelles, such as the nucleus and vacuoles, are easily visible with the light microscope. They were among the first biological discoveries made after the invention of the microscope.

Not all eukaryotic cells have each of the organelles listed below. Exceptional organisms have cells which do not include some organelles that might otherwise be considered universal to eukaryotes (such as mitochondria).^[22] There are also occasional exceptions to the number of membranes surrounding organelles, listed in the tables below (e.g., some that are listed as double-membrane are sometimes found with single or triple membranes). In addition, the number of individual organelles of each type found in a given cell varies depending upon the function of that cell.

Major eukaryotic organelles

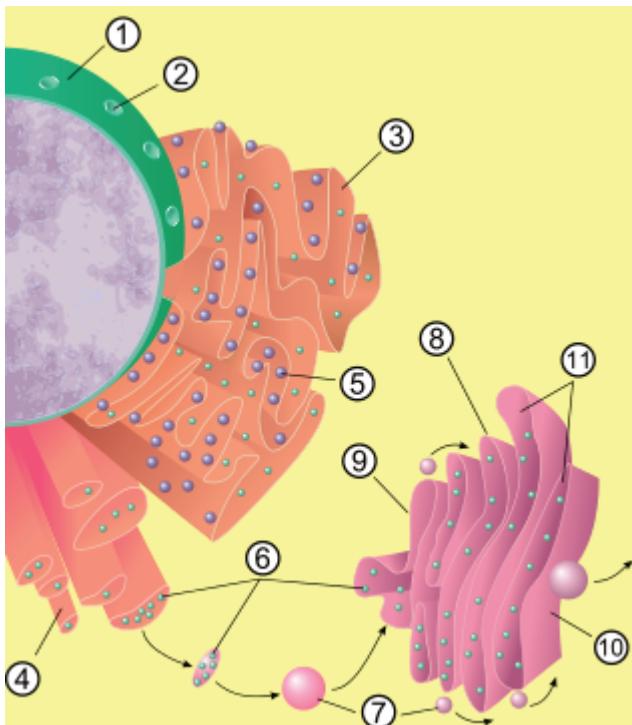
Organelle	Main function
Chloroplast (plastid)	 <p>photosynthesis</p>

The simplified internal structure of a chloroplast

Endoplasmic reticulum

translation and folding of new proteins (rough endoplasmic reticulum), expression of lipids (smooth endoplasmic reticulum)

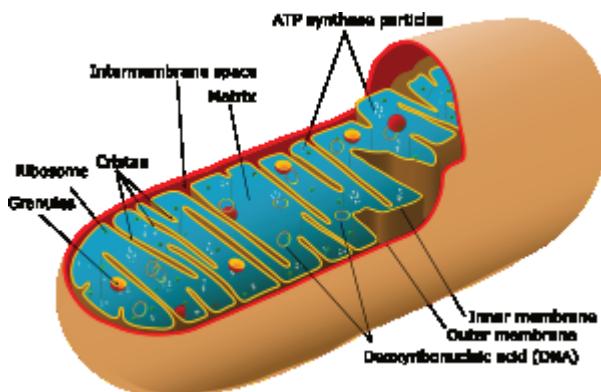
Golgi apparatus



sorting and
modification
of proteins

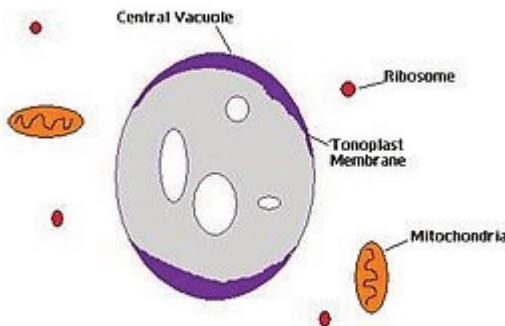
Diagram of secretory process from endoplasmic reticulum (orange) to Golgi apparatus (pink). 1. Nuclear membrane; 2. Nuclear pore; 3. Rough endoplasmic reticulum (RER); 4. Smooth endoplasmic reticulum (SER); 5. Ribosome attached to RER; 6. Macromolecules; 7. Transport vesicles; 8. Golgi apparatus; 9. Cis face of Golgi apparatus; 10. Trans face of Golgi apparatus; 11. Cisternae of lipids

Mitochondria



energy production (house). Mitochondria are self-replicating organelles that occur in various numbers, shapes, and sizes in the cytoplasm of all eukaryotic cells.

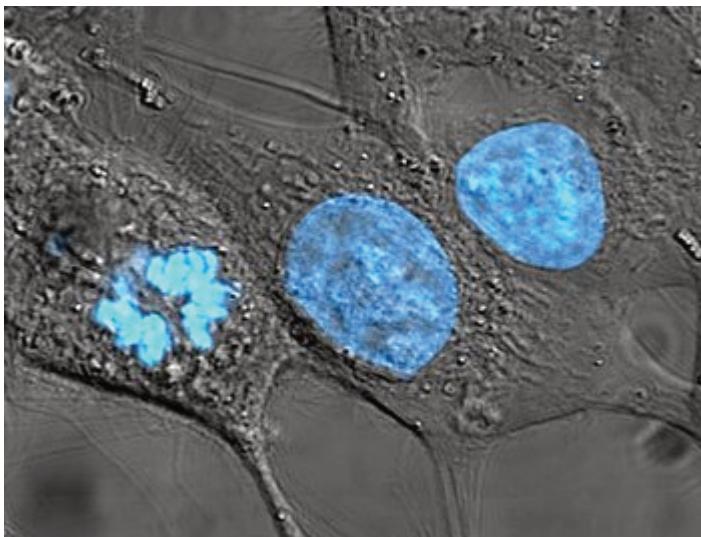
Vacuole



storage, helps maintain homeostasis

The central vacuole within the cytoplasm of a plant cell

Nucleus



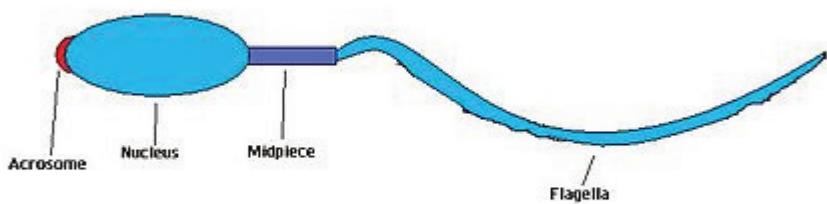
It houses the cell's chromosomes, and is the place where almost all DNA replication, RNA transcription take place

HeLa cells stained for DNA with the Blue Hoechst dye. The central and rightmost cell are in interphase, thus their entire nuclei are labeled. On the left, a cell is going through mitosis and its DNA has condensed ready for division.

Mitochondria and chloroplasts, which have double-membranes and their own DNA, are believed to have originated from incompletely consumed or invading prokaryotic organisms, which were adopted as a part of the invaded cell. This idea is supported in the Endosymbiotic theory.

Organelle/Macromolecule

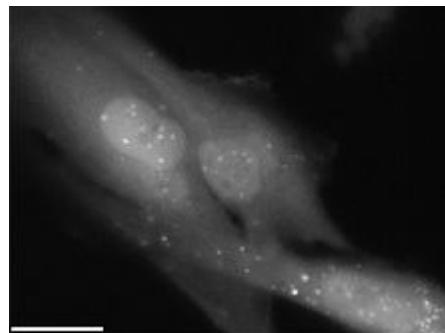
Acrosome



Spermazoa with Acrosome colored red

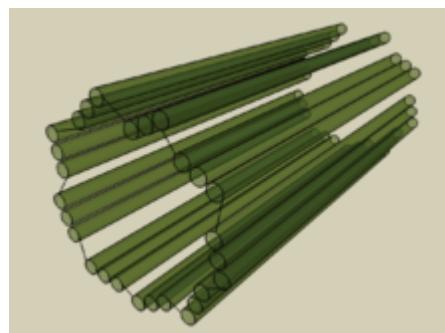
A picture of a spermazoa cell with its acrosome colored in red

Autophagosome



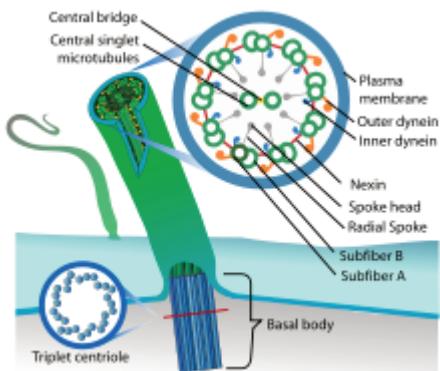
Autophagosomes labeled by a fluorescent marker.

Centriole

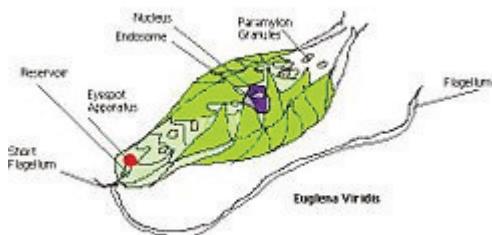


Three-dimensional view of a centriole

Cilium



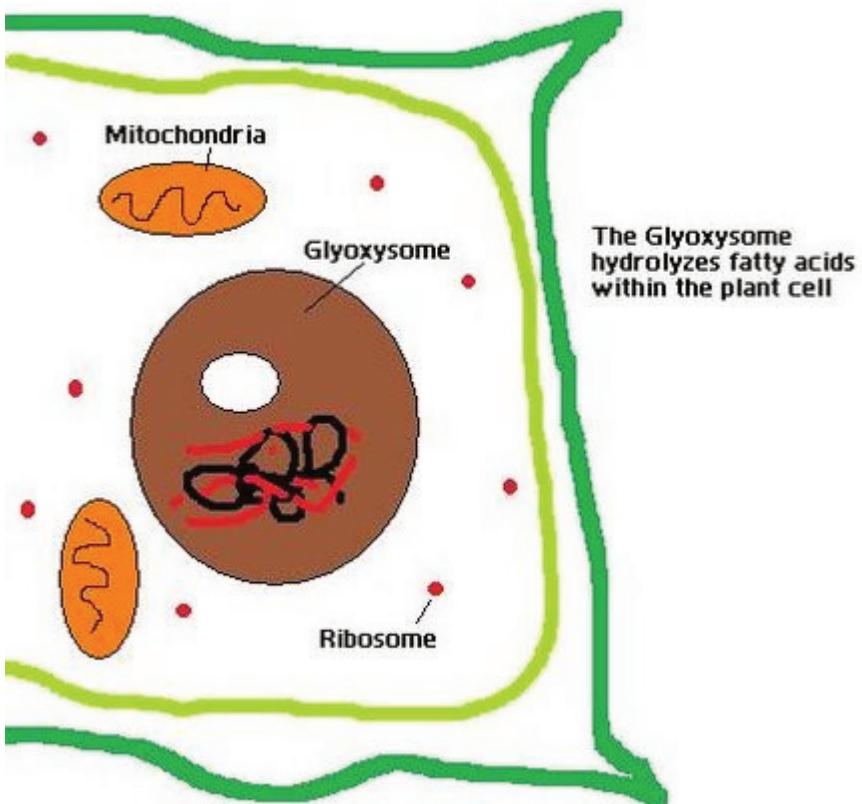
Eukaryotic motile cilium



Euglena Viridis has an eyespot

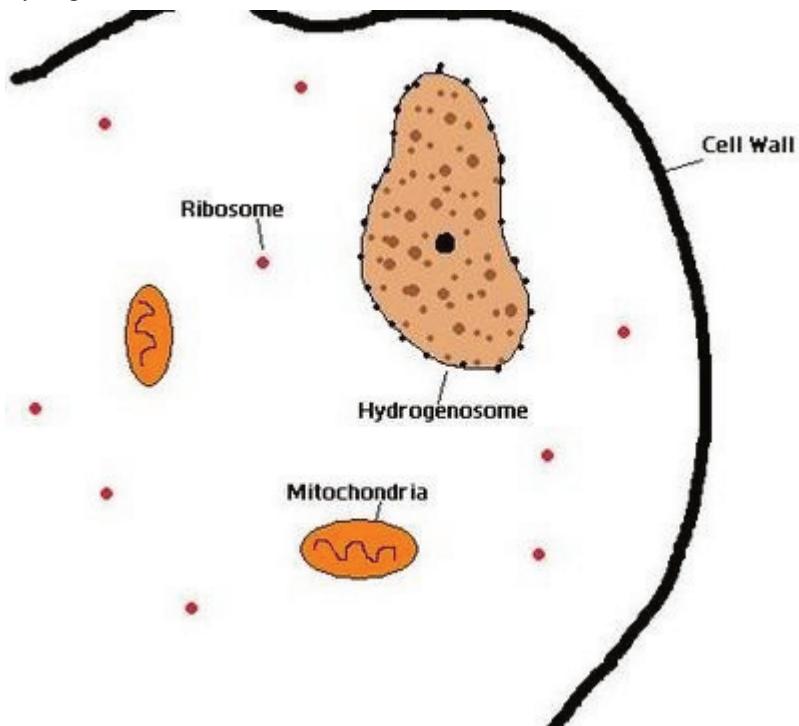
Glycosome

Glyoxysome



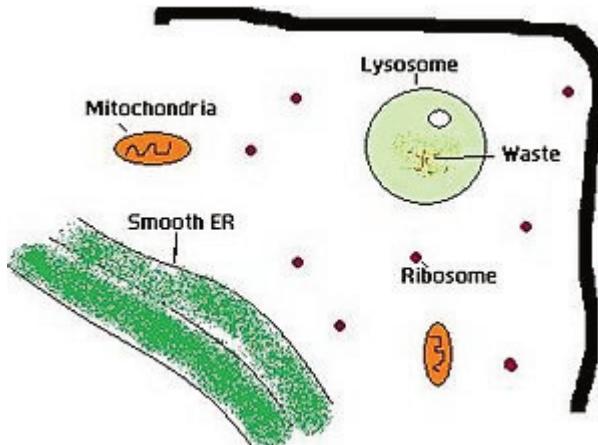
A picture of a Glyoxysome hydrolyzing fatty acids within a plant cell

Hydrogenosome

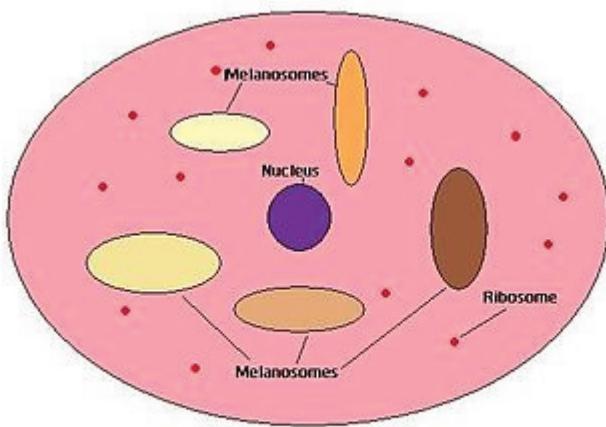


A picture of a hydrogenosome within a eukaryotic cell

Lysosome



Melosome

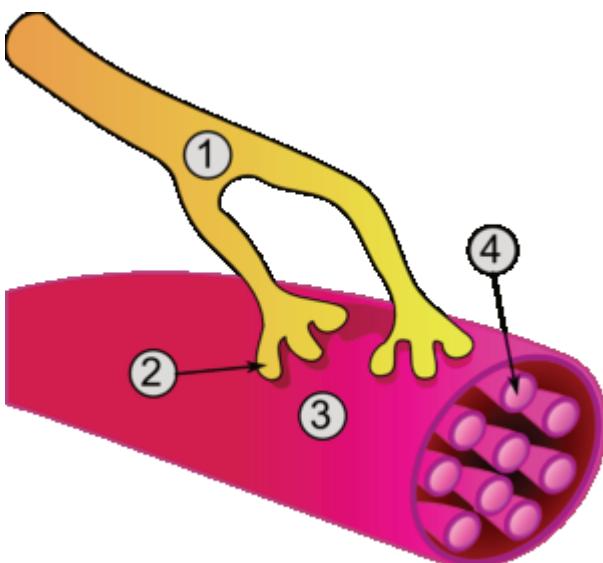


Animal Cell containing Melosomes with various amounts of melanin

Melosomes with varying amounts of melanin in an animal cell

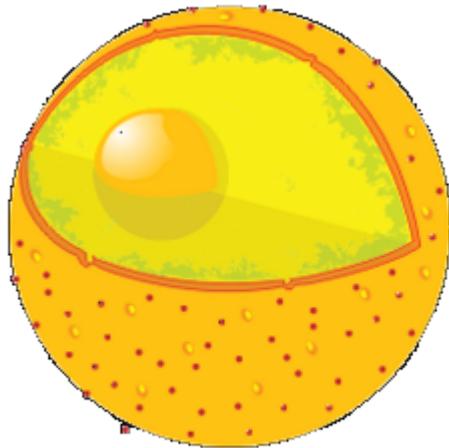
Mitosome

Myofibril



1. Axon
2. Neuromuscular junction
3. Muscle fiber
4. Myofibril

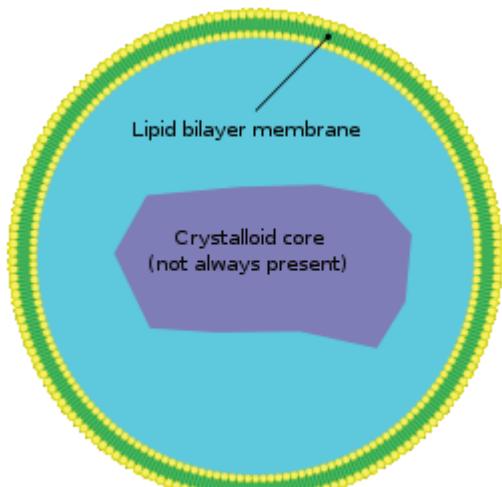
Nucleolus



A nucleolus within a nucleus

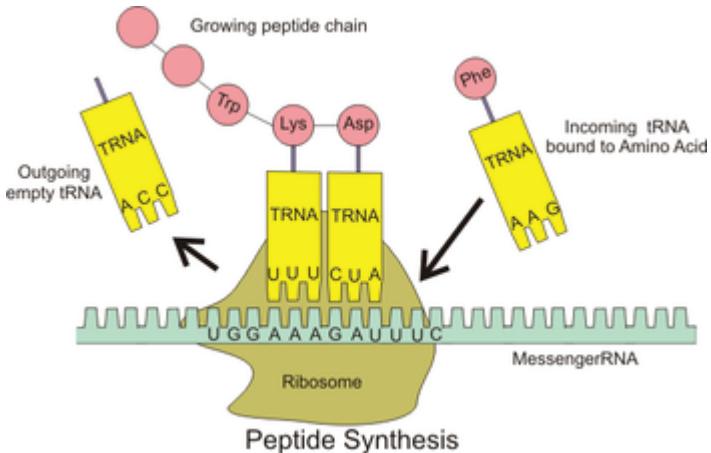
Parenthesome

Peroxisome



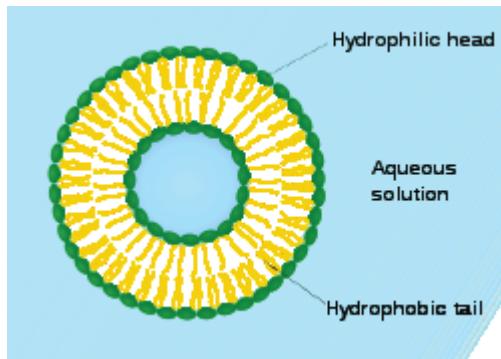
Basic structure of a peroxisome

Ribosome



Ribosomes read the sequence of messenger RNAs and assemble proteins out of amino acids bound to transfer RNAs

vesicle



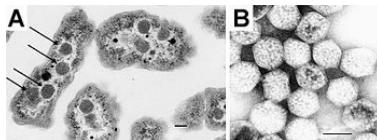
Scheme of a simple vesicle (liposome).

Prokaryotic organelles

Prokaryotes are not as structurally complex as eukaryotes, and were once thought not to have any internal structures enclosed by lipid membranes. In the past, they were often viewed as

having little internal organization; but, slowly, details are emerging about prokaryotic internal structures. An early false turn was the idea developed in the 1970s that bacteria might contain membrane folds termed mesosomes, but these were later shown to be artifacts produced by the chemicals used to prepare the cells for electron microscopy.^[25]

However, more recent research has revealed that at least some prokaryotes have microcompartments such as carboxysomes. These subcellular compartments are 100 – 200 nm in diameter and are enclosed by a shell of proteins.^[26] Even more striking is the description of membrane-bound magnetosomes in bacteria,^{[27][28]} as well as the nucleus-like structures of the Planctomycetes that are surrounded by lipid membranes.^[29]



(A) Electron micrograph of *Halothiobacillus neapolitanus* cells, arrows highlight carboxysomes. (B) Image of intact carboxysomes isolated from *H. neapolitanus*. Scale bars are 100 nm.^[24]

Prokaryotic organelles and cell components

Organelle/ Macromolecule	Main function	Structure	Organisms
Carboxysome	carbon fixation	protein-shell compartment	some bacteria
Chlorosome	photosynthesis	light harvesting complex	green sulfur bacteria
Flagellum	movement in external medium	protein filament	some prokaryotes and eukaryotes
Magnetosome	magnetic orientation	inorganic crystal, lipid membrane	magnetotactic bacteria
Nucleoid	DNA maintenance, transcription to RNA	DNA-protein	prokaryotes
Plasmid	DNA exchange	circular DNA	some bacteria
Ribosome	translation of RNA into proteins	RNA-protein	eukaryotes, prokaryotes
Thylakoid	photosynthesis	photosystem proteins and pigments	mostly cyanobacteria

Macromolecules which are present in the cell membrane

Cell membranes contain a variety of biological molecules, notably lipids and proteins. Material is incorporated into the membrane, or deleted from it, by a variety of mechanisms: Fusion of intracellular vesicles with the membrane (endocytosis) not only excretes the contents of the vesicle but also incorporates the vesicle membrane's components into the cell membrane. The membrane may form blebs around extracellular material that pinch off to become vesicles (exocytosis). If a membrane is continuous with a tubular structure made of membrane material, then material from the tube can be drawn into the membrane continuously. Although the

concentration of membrane components in the aqueous phase is low (stable membrane components have low solubility in water), there is an exchange of molecules between the lipid and aqueous phases.

Lipids

The cell membrane consists of three classes of amphipathic lipids: phospholipids, glycolipids, and cholesterol. The amount of each depends upon the type of cell, but in the majority of cases phospholipids are the most abundant. In RBC studies, 30% of the plasma membrane is lipid. The fatty chains in phospholipids and glycolipids usually contain an even number of carbon atoms, typically between 16 and 20. The 16- and 18-carbon fatty acids are the most common. Fatty acids may be saturated or unsaturated, with the configuration of the double bonds nearly always cis. The length and the degree of unsaturation of fatty acid chains have a profound effect on membrane fluidity as unsaturated lipids create a kink, preventing the fatty acids from packing together as tightly, thus decreasing the melting temperature (increasing the fluidity) of the membrane. The ability of some organisms to regulate the fluidity of their cell membranes by altering lipid composition is called homeoviscous adaptation. The entire membrane is held together via non-covalent interaction of hydrophobic tails, however the structure is quite fluid and not fixed rigidly in place. Under physiological conditions phospholipid molecules in the cell membrane are in the liquid crystalline state. It means the lipid molecules are free to diffuse and exhibit rapid lateral diffusion along the layer in which they are present. However, the exchange of phospholipid molecules between intracellular and extracellular leaflets of the bilayer is a very slow process. Lipid rafts and caveolae are examples of cholesterol-enriched microdomains in the cell membrane. In animal cells cholesterol is normally found dispersed in varying degrees throughout cell membranes, in the irregular spaces between the hydrophobic tails of the membrane lipids, where it confers a stiffening and strengthening effect on the membrane. Lipid vesicles or liposomes are circular pockets that are

enclosed by a lipid bilayer. These structures are used in laboratories to study the effects of chemicals in cells by delivering these chemicals directly to the cell, as well as getting more insight into cell membrane permeability. Lipid vesicles and liposomes are formed by first suspending a lipid in an aqueous solution then agitating the mixture through sonication, resulting in a uniformly circular vesicle. By measuring the rate of efflux from that of the inside of the vesicle to the ambient solution, allows researcher to better understand membrane permeability. Vesicles can be formed with molecules and ions inside the vesicle by forming the vesicle with the desired molecule or ion present in the solution. Proteins can also be embedded into the membrane through solubilizing the desired proteins in the presence of detergents and attaching them to the phospholipids in which the liposome is formed. These provide researchers with a tool to examine various membrane protein functions.

Carbohydrates

Plasma membranes also contain carbohydrates, predominantly glycoproteins, but with some glycolipids (cerebrosides and gangliosides). For the most part, no glycosylation occurs on membranes within the cell; rather generally glycosylation occurs on the extracellular surface of the plasma membrane. The glycocalyx is an important feature in all cells, especially epithelia with microvilli. Recent data suggest the glycocalyx participates in cell adhesion, lymphocyte homing, and many others. The penultimate sugar is galactose and the terminal sugar is sialic acid, as the sugar backbone is modified in the golgi apparatus. Sialic acid carries a negative charge, providing an external barrier to charged particles.

Proteins

Proteins within the membrane are key to the functioning of the overall membrane. These proteins mainly transport chemicals and information across the membrane. Every membrane has a varying degree of protein content. Proteins can be in the form of peripheral or integral. The cell membrane plays host to a large amount of protein that is responsible for its various activities. The amount of

protein differs between species and according to function, however the typical amount in a cell membrane is 50%. These proteins are undoubtedly important to a cell: Approximately a third of the genes in yeast code specifically for them, and this number is even higher in multicellular organisms. The cell membrane, being exposed to the outside environment, is an important site of cell-cell communication. As such, a large variety of protein receptors and identification proteins, such as antigens, are present on the surface of the membrane. Functions of membrane proteins can also include cell-cell contact, surface recognition, cytoskeleton contact, signaling, enzymatic activity, or transporting substances across the membrane. Most membrane proteins must be inserted in some way into the membrane. For this to occur, an N-terminus “signal sequence” of amino acids directs proteins to the endoplasmic reticulum, which inserts the proteins into a lipid bilayer. Once inserted, the proteins are then transported to their final destination in vesicles, where the vesicle fuses with the target membrane.

Facts to be remembered

Some theorists suggest that the atmosphere of the early Earth may have been chemically reducing in nature, composed primarily of methane (CH_4), ammonia (NH_3), water (H_2O), hydrogen sulfide (H_2S), carbon dioxide (CO_2) or carbon monoxide (CO), and phosphate (PO_4^{3-}), with molecular oxygen (O_2) and ozone (O_3) either rare or absent.

The sequence of chemical events that led to the first nucleic acids is not known.

In such a reducing atmosphere, electrical activity can catalyze the creation of certain basic small molecules (monomers) of life, such as amino acids. This was demonstrated in the Miller–Urey experiment by Stanley L. Miller and Harold C. Urey in 1953.

Phospholipids (of an appropriate length) can spontaneously form lipid bilayers, a basic component of the cell membrane.

The polymerization of nucleotides into random RNA molecules might have resulted in self-replicating ribozymes.

Synthesized proteins might then outcompete ribozymes in catalytic ability, and therefore become the dominant biopolymer, relegating nucleic acids to their modern use, predominantly as a carrier of genomic information.

1632–1723: Antonie van Leeuwenhoek teaches himself to grind lenses, builds a microscope and draws protozoa, such as Vorticella from rain water, and bacteria from his own mouth.

1665: Robert Hooke discovers cells in cork, then in living plant tissue using an early microscope.

1839: Theodor Schwann and Matthias Jakob Schleiden elucidate the principle that plants and animals are made of cells, concluding that cells are a common unit of structure and development, and thus founding the cell theory.

The belief that life forms can occur spontaneously (generatio spontanea) is contradicted by Louis Pasteur (1822–1895) (although Francesco Redi had performed an experiment in 1668 that suggested the same conclusion).

1855: Rudolf Virchow states that cells always emerge from cell divisions (omnis cellula ex cellula).

1931: Ernst Ruska builds first transmission electron microscope (TEM) at the University of Berlin. By 1935, he has built an EM with twice the resolution of a light microscope, revealing previously unresolvable organelles.

1953: Watson and Crick made their first announcement on the double-helix structure for DNA on February 28.

1981: Lynn Margulis published Symbiosis in Cell Evolution detailing the endosymbiotic theory.

Evidence which support endosymbiotic theory

Mitochondria and plastids are formed only through a process similar to binary fission. In some algae, such as Euglena, the plastids can be destroyed by certain chemicals or prolonged absence of light

without otherwise affecting the cell. In such a case, the plastids will not regenerate.

They are surrounded by two or more membranes, and the innermost of these shows differences in composition from the other membranes of the cell. They are composed of a peptidoglycan cell wall characteristic of a bacterial cell.

Both mitochondria and plastids contain DNA that is different from that of the cell nucleus and that is similar to that of bacteria (in being circular in shape and in its size). DNA sequence analysis and phylogenetic estimates suggest that nuclear DNA contains genes that probably came from plastids.

These organelles' ribosomes are like those found in bacteria (70S).

Proteins of organelle origin, like those of bacteria, use N-formylmethionine as the initiating amino acid.

Much of the internal structure and biochemistry of plastids, for instance the presence of thylakoids and particular chlorophylls, is very similar to that of cyanobacteria.

Phylogenetic estimates constructed with bacteria, plastids, and eukaryotic genomes also suggest that plastids are most closely related to cyanobacteria. Mitochondria have several enzymes and transport systems similar to those of bacteria.

Some proteins encoded in the nucleus are transported to the organelle, and both mitochondria and plastids have small genomes compared to bacteria. This is consistent with an increased dependence on the eukaryotic host after forming an endosymbiosis. Most genes on the organellar genomes have been lost or moved to the nucleus. Most genes needed for mitochondrial and plastid function are located in the nucleus. Many originate from the bacterial endosymbiont. Plastids are present in very different groups of protists, some of which are closely related to forms lacking plastids. This suggests that if chloroplasts originated de novo, they did so multiple times, in which case their close similarity to each other is difficult to explain.

Many of these protists contain "primary" plastids that have not yet been acquired from other plastid-containing eukaryotes. Among

eukaryotes that acquired their plastids directly from bacteria (known as Primoplantae), the glaucophyte algae have chloroplasts that strongly resemble cyanobacteria. In particular, they have a peptidoglycan cell wall between the two membranes. Mitochondria and plastids are similar in size to bacteria.

Question time

1. Cell with large round size has more chance to survive as compare to thin cell under desiccation, why?
 1. a. because of thin membrane
 2. b. because of thick membrane
 3. c. none of the above
 4. d. high surface to volume ratio
2. What are the difference between Oligosaccharides and Polysaccharides?
3. Which cell did evolve first prokaryote or eukaryote?
4. What is endosymbiotic theory? could it happen present time also if yes then how if not then why not?
5. Which reaction is not possible in biological system?
 1. a.DNA-RNA-protein
 2. b.glucose-aminoacid-protein
 3. c.protein-RNA-DNA
 4. d.RNA-DNA-Protein
6. What is the difference between prokaryote and eukaryote?
7. How are the plant cells different from animal cell?
8. Which kind of gases were present before the origin of life?

References

1. ↑ Jump up to: **a b** “Right-handed amino acids were left behind”.

- New Scientist (Reed Business Information Ltd) (2554): pp. 18. 2006-06-02. <http://www.newscientist.com/channel/life/mg19025545.200-right-handed-amino-acids-were-left-behind.html>. Retrieved 2008-07-09.
2. ↑ Kojo, Shosuke; Hiromi Uchino, Mayu Yoshimura and Kyoko Tanaka (October 2004). "Racemic D,L-asparagine causes enantiomeric excess of other coexisting racemic D,L-amino acids during recrystallization: a hypothesis accounting for the origin of L-amino acids in the biosphere". *Chemical Communications* (19): 2146–2147. doi:10.1039/b409941a. PMID 15467844.
 3. ↑ "MICR 425: PHYSIOLOGY & BIOCHEMISTRY of MICROORGANISMS: The Origin of Life". SIUC / College of Science. <http://www.science.siu.edu/microbiology/micr425/425Notes/14-OriginLife.html>. Retrieved 2005-12-17.
 4. ↑ "Early Earth atmosphere favorable to life: study". University of Waterloo. <http://newsrelease.uwaterloo.ca/news.php?id=4348>. Retrieved 2005-12-17.
 5. ↑ Fitzpatrick, Tony (2005). "Calculations favor reducing atmosphere for early earth – Was Miller–Urey experiment correct?". Washington University in St. Louis. <http://news-info.wustl.edu/news/page/normal/5513.html>. Retrieved 2005-12-17.
 6. ↑ Thompson WR, Murray BG, Khare BN, Sagan C (December 1987). "Coloration and darkening of methane clathrate and other ices by charged particle irradiation: applications to the outer solar system". *Journal of geophysical research* **92** (A13): 14933–47. doi:10.1029/JA092iA13p14933. PMID 11542127.
 7. ↑ Hill HG, Nuth JA (2003). "The catalytic potential of cosmic dust: implications for prebiotic chemistry in the solar nebula and other protoplanetary systems". *Astrobiology* **3** (2): 291–304. doi:10.1089/153110703769016389. PMID 14577878.

8. ↑ Balm SP, Hare J.P., Kroto HW (1991). "The analysis of comet mass spectrometric data". *Space Science Reviews* **56**: 185–9. doi:10.1007/BF00178408.
9. ↑ Bada, Jeffrey L. (2000). "Stanley Miller's 70th Birthday" (PDF). *Origins of Life and Evolution of the Biosphere* (Netherlands: Kluwer Academic Publishers) **30**: 107–12. doi:10.1023/A:1006746205180. <http://www.issol.org/miller/70thB-Day.pdf>.
10. ↑ Miller, Stanley L. (May 1953). "Production of Amino Acids Under Possible Primitive Earth Conditions" (PDF). *Science* **117** (3046): 528. doi:10.1126/science.117.3046.528. PMID 13056598. http://www.abenteuer-universum.de/pdf/miller_1953.pdf.
11. ↑ Miller, Stanley L.; Harold C. Urey (July 1959). "Organic Compound Synthesis on the Primitive Earth". *Science* **130** (3370): 245. doi:10.1126/science.130.3370.245. PMID 13668555. Miller states that he made "A more complete analysis of the products" in the 1953 experiment, listing additional results.
12. ↑ A. Lazcano, J. L. Bada (June 2004). "The 1953 Stanley L. Miller Experiment: Fifty Years of Prebiotic Organic Chemistry". *Origins of Life and Evolution of Biospheres* **33** (3): 235–242. doi:10.1023/A:1024807125069. PMID 14515862.
13. ↑ "EXOBIOLOGY: An Interview with Stanley L. Miller". [Accessexcellence.org](http://www.accessexcellence.org/WN/NM/miller.php). <http://www.accessexcellence.org/WN/NM/miller.php>. Retrieved 2009-08-20.
14. ↑ Dreifus, Claudia (2010-05-17). "A Conversation With Jeffrey L. Bada: A Marine Chemist Studies How Life Began". [nytimes.com](http://www.nytimes.com/2010/05/18/science/18conv.html). <http://www.nytimes.com/2010/05/18/science/18conv.html>.
15. ↑ http://books.nap.edu/openbook.php?record_id=11860&page=85 Exploring Organic Environments in the Solar System (2007)
16. ↑ Oró J, Kimball AP (August 1961). "Synthesis of purines under possible primitive earth conditions. I. Adenine from hydrogen cyanide". *Archives of biochemistry and biophysics* **94**: 217–27. doi:10.1016/0003-9861(61)90033-9. PMID 13731263.
17. ↑ Oró J, Kamat SS (April 1961). "Amino-acid synthesis from

- hydrogen cyanide under possible primitive earth conditions". *Nature* **190**: 442–3. doi:10.1038/190442a0. PMID 13731262.
- 18. ↑ Oró J (1967). Fox SW. ed. *Origins of Prebiological Systems and of Their Molecular Matrices*. New York Academic Press. pp. 137.
 - 19. ↑ Wilde, Kenneth A.; Bruno J. Zwolinski and Ransom B. Parlin (July 1953). "The Reaction Occurring in CO₂, O₂ Mixtures in a High-Frequency Electric Arc". *Science* **118** (3054): 43–44. doi:10.1126/science.118.3054.43-a. PMID 13076175. <http://www.sciencemag.org/cgi/content/citation/118/3054/43-a>. Retrieved 2008-07-09.
 - 20. ↑ Synthesis of organic compounds from carbon monoxide and water by UV photolysis
 - 21. ↑ Fox, Douglas (2007-03-28). "Primordial Soup's On: Scientists Repeat Evolution's Most Famous Experiment". *Scientific American* (Scientific American Inc.). <http://www.sciam.com/article.cfm?id=primordial-soup-urey-miller-evolution-experiment-repeated>. Retrieved 2008-07-09.
 - 22. ↑ Fahey RC, Newton GL, Arrack B, Overdank-Bogart T, Baley S (1984). "Entamoeba histolytica: a eukaryote without glutathione metabolism". *Science* **224** (4644): 70–72. doi:10.1126/science.6322306. PMID 6322306.
 - 23. ↑ Badano, Jose L; Norimasa Mitsuma, Phil L. Beales, Nicholas Katsanis (September 2006). "The Ciliopathies : An Emerging Class of Human Genetic Disorders". *Annual Review of Genomics and Human Genetics* **7**: 125–148. doi:10.1146/annurev.genom.7.080505.115610. PMID 16722803. <http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.genom.7.080505.115610>. Retrieved 2008-06-15.
 - 24. ↑ Tsai Y, Sawaya MR, Cannon GC, Cai F, Williams EB, Heinhorst S, Kerfeld CA, Yeates TO (Jun 2007). "Structural analysis of CsoS1A and the protein shell of the Halothiobacillus neapolitanus carboxysome." (Free full text). *PLoS biology* **5** (6): e144. doi:10.1371/journal.pbio.0050144. PMID 17518518. PMC 1872035. <http://biology.plosjournals.org/>

- perlServ/?request=get-document&doi=10.1371/journal.pbio.0050144.
- 25. ↑ Ryter A (1988). "Contribution of new cryomethods to a better knowledge of bacterial anatomy". *Ann. Inst. Pasteur Microbiol.* **139** (1): 33–44. doi:10.1016/0769-2609(88)90095-6. PMID 3289587.
 - 26. ↑ Kerfeld CA, Sawaya MR, Tanaka S, Nguyen CV, Phillips M, Beeby M, Yeates TO (August 2005). "Protein structures forming the shell of primitive bacterial organelles.". *Science* **309** (5736): 936–8. doi:10.1126/science.1113397. PMID 16081736.
 - 27. ↑ Komeili A, Li Z, Newman DK, Jensen GJ (2006). "Magnetosomes are cell membrane invaginations organized by the actin-like protein MamK". *Science* **311** (5758): 242–5. doi:10.1126/science.1123231. PMID 16373532.
 - 28. ↑ Scheffel A, Gruska M, Faivre D, Linaroudis A, Plitzko JM, Schüler D (2006). "An acidic protein aligns magnetosomes along a filamentous structure in magnetotactic bacteria". *Nature* **440** (7080): 110–4. doi:10.1038/nature04382. PMID 16299495.
 - 29. ↑ Fuerst JA (2005). "Intracellular compartmentation in planctomycetes". *Annu. Rev. Microbiol.* **59**: 299–328. doi:10.1146/annurev.micro.59.030804.121258. PMID 15910279.

IO.

The nucleus was the first organelle to be discovered. The probably oldest preserved drawing dates back to the early microscopist Antonie van Leeuwenhoek (1632 – 1723). He observed a “Lumen,” the nucleus, in the red blood cells of salmon. Unlike mammalian red blood cells, those of other vertebrates still possess nuclei. The nucleus was also described by Franz Bauer in 1804 and in 1831 by Scottish botanist **Robert Brown** in a talk at the Linnean Society of London. Brown was studying orchids under microscope when he observed an opaque area, which he called the areola or nucleus, in the cells of the flower’s outer layer. He did not suggest a potential function.

In 1838, Matthias Schleiden proposed that the nucleus plays a role in generating cells, thus he introduced the name “**Cytoblast**” (cell builder). He believed that he had observed new cells assembling around “cytoblasts”. Franz Meyen was a strong opponent of this view, having already described cells multiplying by division and believing that many cells would have no nuclei. The idea that cells can be generated de novo, by the “cytoblast” or otherwise, contradicted work by Robert Remak (1852) and Rudolf Virchow (1855) who decisively propagated the new paradigm that cells are generated solely by cells (“**Omnis cellula e cellula**”).

The function of the nucleus remained unclear. Between 1876 and 1878, Oscar Hertwig published several studies on the fertilization of sea urchin eggs, showing that the nucleus of the sperm enters the oocyte and fuses with its nucleus. This was the first time it was suggested that an individual develops from a (single) nucleated cell. This was in contradiction to Ernst Haeckel’s theory that the complete phylogeny of a species would be repeated during embryonic development, including generation of the first nucleated cell from a “Monerula”, a structureless mass of primordial mucus (“Urschleim”). Therefore, the necessity of the sperm nucleus for

fertilization was discussed for quite some time. However, Hertwig confirmed his observation in other animal groups, e.g., amphibians and molluscs. **Eduard Strasburger** produced the same results for plants (1884). This paved the way to assign the nucleus an important role in heredity. In 1873, August Weismann postulated the equivalence of the maternal and paternal germ cells for heredity. The function of the nucleus as carrier of genetic information became clear only later, after mitosis was discovered and the Mendelian rules were rediscovered at the beginning of the 20th century; the chromosome theory of heredity was developed.^[1]

Contents

- 1 Nucleus
- 2 Components of NUCLEUS
 - 2.1 The nucleoskeleton
 - 2.2 Chromosomes
 - 2.3 Nucleolus
 - 2.4 Splicing speckles or SC35 speckles
 - 2.5 Cajal bodies or coiled bodies
- 3 Nuclear import and export
 - 3.1 Nuclear localization signal (NLS)
 - 3.2 Nuclear export signal (NES)
 - 3.2.1 Role of Ran GTPase in nuclear transport during interphase
- 4 Facts to be remembered
- 5 Question time
- 6 References

Nucleus

The nucleus (**pl. nuclei**; from Latin *nucleus* or *nuculeus*, meaning kernel) is a membrane enclosed organelle found in eukaryotic cells. It contains most of the cell's genetic material (DNA), organized as multiple long linear DNA molecules in complex with a large variety of proteins, such as histones, to form chromosomes. The genes within these chromosomes are the cell's nuclear genome. The function of the nucleus is to maintain the integrity of these genes and to control the activities of the cell by regulating gene expression – the nucleus is therefore the control center of the cell. The main structures making up the nucleus are the nuclear envelope, a double membrane that encloses the entire organelle and separates its contents from the cellular cytoplasm, and the nuclear lamina, a meshwork within the nucleus that adds mechanical support, much like the cytoskeleton supports the cell as a whole. Because the nuclear membrane is impermeable to most molecules, nuclear pores are required to allow movement of molecules across the envelope. These pores cross both of the membranes, providing a channel that allows free movement of small molecules and ions. The movement of larger molecules such as proteins is carefully controlled, and requires active transport regulated by carrier proteins. Nuclear transport is crucial to cell function, as movement through the pores is required for both gene expression and chromosomal maintenance. Although the interior of the nucleus does not contain any membrane-bound subcompartments, its

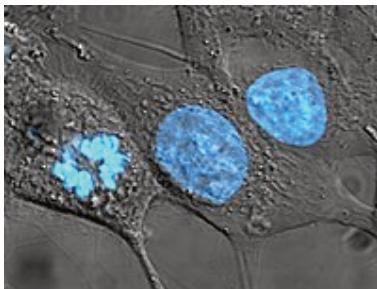


Fig. 1 HeLa cells stained for DNA with the Blue Hoechst dye. The central and rightmost cell are in interphase, thus their entire nuclei are labeled. On the left a cell is going through mitosis and its DNA has condensed ready for division.

contents are not uniform, and a number of subnuclear bodies exist, made up of unique proteins, RNA molecules, and particular parts of the chromosomes. The best known of these is the nucleolus, which is mainly involved in the assembly of ribosomes. After being produced in the nucleolus, ribosomes are exported to the cytoplasm where they translate mRNA.^[2]

Components of NUCLEUS

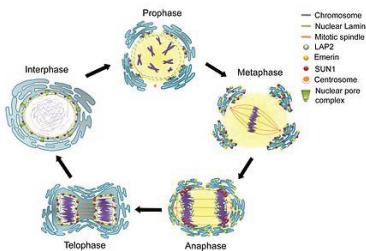
The nucleoskeleton

The nucleoskeletal framework that remains insoluble after treatment of nuclei with non-ionic detergents, followed by nuclease treatment and high salt extraction to remove chromatin and soluble proteins, is generally termed the nuclear matrix. This nucleoskeleton consists of two parts namely, the nuclear lamina and a network of intricately structured fibres connected to the lamina and distributed throughout the nuclear volume. These highly structured fibre assemblies have been shown to be attached to an underlying network of 10 nm core filaments . The nuclear lamins A, B and C are the major structural components of the peripheral lamina. Additional proteins that connect the lamina to the nuclear envelope and the heterochromatin are clustered at the nuclear periphery. Other nuclear matrix-associated components like hnRNP complexes, newly transcribed full-length mRNA, RNA polymerase I and II, various transcription factors and spliceosomal complexes are positioned on the underlying network of branched 10 nm filaments that is connected to the nuclear lamina . These associations might be dynamic and allow for considerable plasticity in nuclear architecture and function. Furthermore, numerous studies have suggested the presence of an organizing structure such as the nuclear matrix to coordinate the spatial regulation of DNA

synthesis. The major candidate proteins that are likely to comprise the nucleoskeleton are lamins and nuclear actin.

1. Nuclear envelope:

The nuclear envelope (NE) (also known as the perinuclear envelope, nuclear membrane, nucleolemma or karyotheca) is a double lipid bilayer that encloses the genetic material in eukaryotic cells. The nuclear envelope also serves as the physical barrier, separating the contents of the nucleus (DNA in particular) from the cytosol (cytoplasm). Many nuclear pores are inserted in the nuclear envelope, which facilitate and regulate the exchange of materials (proteins such as transcription factors, and RNA) between the nucleus and the cytoplasm. Nuclear membranes is composed of a lipid bilayer. The outer membrane is continuous with the rough endoplasmic reticulum while the inner nuclear membrane is the primary residence of several inner nuclear membrane proteins. **The outer and inner nuclear membrane are fused at the site of nuclear pore complexes.** The structure of the membrane also consists of ribosomes. The space between the two membranes that make up the nuclear envelope is called the perinuclear space (also called the perinuclear cisterna, NE Lumen), and is usually about 20 – 40 nm wide. The inner nuclear membrane is connected to the nuclear lamina.^[4] **Nuclear envelope breakdown during mitosis**



Nuclear envelope breakdown and reassembly in mitosis. At the end of G2, the activation of cyclin-dependent kinases, including CDK1, triggers entry into mitotic prophase. The nuclear membrane breaks down, and the NE-associate proteins either translocate to kinetochores, distribute with the fragmented ER networks, or dissolve in the cytoplasm. During NE reassembly in anaphase, SUN1 and LAP2 first appear around the condensed chromatin, though at different regions. The nuclear lamins then join the nuclear periphery in telophase.
This figure illustrates the important roles played by the NE and the nuclear lamina in normal mitosis. Chi et al. Journal of Biomedical Science 2009.^[3]

envelope breakdown and reassembly in mitosis. At the end of G2, the activation of cyclin-dependent kinases, including CDK1, triggers entry into mitotic prophase. The nuclear membrane breaks down, and the NE-associate proteins either translocate to kinetochores, distribute with the fragmented ER networks, or dissolve in the cytoplasm. During NE reassembly in anaphase, SUN1 and LAP2 first appear around the condensed chromatin, though at different regions. The nuclear lamins then join the nuclear periphery in telophase. This figure illustrates the important roles played by the NE and the nuclear lamina in normal mitosis. Chi et al. Journal of Biomedical Science 2009.[1] During prophase in mitosis, the chromatids begin condensing to form chromosomes, and the nuclear envelope breaks down and is retracted into the mitotic endoplasmic reticulum. At metaphase, the nuclear envelope has been completely disassembled and absorbed by the ER allowing the chromosomes to be put together by spindle fibers attached to each chromosome at the kinetochore. Other eukaryotes such as yeast undergo closed mitosis, where the chromosomes segregate within the nuclear envelope, which then buds as the three daughter cells divide. In the process of mitosis, the nuclear envelope is degraded during prometaphase. Without this step, the nuclear material would be unable to separate into two nuclei, and by extension, two cells. At the end of anaphase however, the chromosomes are now separated, and each set is in its respective half of the parent cell. In order to protect the genetic material, the nuclear membrane must re-form at this stage. This is done using membrane from the endoplasmic reticulum and proteins called lamins that guide the new envelope. (Alberts) Like some of the other organelles in the cell, the endoplasmic reticulum is composed of a phospholipid bilayer. This thin membrane is very flexible, and portions can be pinched off to form new organelles, vesicles, or in this case, nuclei. In 2007, researchers discovered that pieces of the ER break off and merge together to form the nuclear envelope. However, there must be a structural element that brings these vesicles together, because the nucleus is much more structurally complicated than the ER. This

element is a layer of lamin between the chromatid and the nuclear envelope itself, known as the lamina. Lamins are long fibrous proteins. In prometaphase, they are phosphorylated (a phosphate group is added), causing the protein to change shape and lose its structural properties. This is what causes the breaking down of the nuclear membrane. However, towards the end of anaphase, the existing lamin is dephosphorylated, and even more is produced. Once the chromosomes have separated, the lamina begins to form again. Sometime at the beginning of mitosis, ends of the endoplasmic reticulum bound to the DNA, using lamin A receptor. (Duband-Goulet and Courvalin) Once the lamina begins to re-form, it forces tubules of ER to form a network across the surface of the chromatid. (Anderson and Hetzer) These tubes eventually are flattened out and merged, forming a solid nuclear membrane. It is not certain what mechanisms cause this to happen. The ER and the LBR eventually detach themselves from the DNA. In a mature cell, the lamina forms a continuous layer just inside the nuclear envelope, and is attached to the nucleus by protein known as emerin.^[5]

2.Nuclear pores:

Nuclear pores are large protein complexes that cross the nuclear envelope, which is the double membrane surrounding the eukaryotic cell nucleus. There are about on average 2000 nuclear pore complexes in the nuclear envelope of a vertebrate cell, but it varies depending on cell type and the stage in the life cycle. The proteins that make up the nuclear pore complex are known as nucleoporins. About half of the nucleoporins typically contain either an alpha solenoid or a beta-propeller fold, or in some cases both as separate structural domains. The other half show structural characteristics typical of “natively unfolded” proteins, i.e. they are highly flexible proteins that lack ordered secondary structure.^[6]

Assembly of the NPC

As the NPC controls access to the genome, it is essential that it exists in large amounts in areas of the cell cycle where plenty

of transcription is necessary. For example, cycling mammalian and yeast cells double the amount of NPC in the nucleus between the G1 and G2 phase of cell Mitosis. And oocytes accumulate large numbers of NPCs to prepare for the rapid mitosis that exists in the early stages of development. Interphase cells must also keep up a level of NPC generation to keep the levels of NPC in the cell constant as some may get damaged. Some cells can even increase the NPC numbers due to increased transcriptional demand.

Theories of assembly So how are these vast proteins complexes assembled? As the immunodepletion of certain protein complexes, such as the Nup 107–160 complex, leads to the formation of poreless nuclei, it seems likely that the Nup complexes are involved in fusing the outer membrane of the nuclear envelope with the inner and not that the fusing of the membrane begins the formation of the pore. There are several ways that this could lead to the formation of the full NPC. One possibility is that as a protein complex it binds to the chromatin. It is then inserted into the double membrane close to the chromatin. This, in turn, leads to the fusing of that membrane. Around this protein complex others eventually bind forming the NPC. This method is possible during every phases of mitosis as the double membrane is present around the chromatin before the membrane fusion proteins complex can insert. Post mitotic cells could form a membrane first with pores being inserted into after formation. Another model for the formation of the NPC is the production of a prepore as a start as opposed to a single protein complex. This prepore would form when several Nup complexes come together and bind to the chromatin. This would have the double membrane form around it in during mitotic reassembly. Possible prepore structures have been observed on chromatin before nuclear envelope(NE) formation using electron microscopy. During the interphase of the cell cycle the formation of the prepore would happen within the nucleus, each component being transported in through existing NPCs. These Nups would bind to an importin, once formed, preventing the assembly of a prepore in the cytoplasm. Once transported into the nucleus Ran GTP would

bind to the importin and cause it to release the cargo. This Nup would be free to from a prepore. The binding of importins has at least been shown to bring Nup 107 and the Nup 153 nucleoporins into the nucleus. NPC assembly is a very rapid process yet defined intermediate states occur which leads to the idea that this assembly occurs in a stepwise fashion. A third possible method of NPC assembly is splitting. This method seems to be tailor made for NPC formation during the interphase. It happens when more protomers are added on to an existing NPC. The eightfold symmetry of the NPC has been shown to have a degree of plasticity and will allow this. Eventually enough protomers will add and allow a new NPC to split off the original. This method of NPC assembly can only happen during the interphase of the cell cycle.

Disassembly of NPC During mitosis the NPC appears to disassemble in stages. Peripheral nucleoporins such as the Nup 153 Nup 98 and Nup 214 disassociate from the NPC. The rest, which can be considered a scaffold proteins remain stable, as cylindrical ring complexes within the nuclear envelope. This disassembly of the NPC peripheral groups is largely thought to be phosphate driven, as several of these nucleoporins are phosphorylated during the stages of mitosis. However, the enzyme involved in the phosphorylation is unknown *in vivo*. In metazoans (which undergo open mitosis) the NE degrades quickly after the loss of the peripheral Nups. The reason for this may be due to the change in the NPC's architecture. This change may make the NPC more permeable to enzymes involved in the degradation of the NE such as cytoplasmic tubulin, as well as allowing the entry of key mitotic regulator proteins.

It was shown, in fungi that undergo closed mitosis (where the nucleus does not degrade), that the change of the permeability barrier of the NE was due to changes with in the NPC and is what allows the entry of mitotic regulators. In *Aspergillus nidulans* the NPC composition appears to be effected by the mitotic kinase NIMA, possibly by phosphorylating the nucleoporins Nup98 and Gle2/Rae1. This remodelling seems to allow the proteins complex cdc2/cyclinB enter the nucleus as well as many other proteins such

as soluble tubulin. The NPC scaffold remains intact throughout the whole closed mitosis. This seems to preserve the integrity of the NE.^[6]

3.Nuclear lamina:

The nuclear lamina is a dense (~30 to 100 nm thick) fibrillar network inside the nucleus of a eukaryotic cell. It is composed of intermediate filaments and membrane associated proteins. Besides providing mechanical support, the nuclear lamina regulates important cellular events such as DNA replication and cell division. Additionally, it participates in chromatin organization and it anchors the nuclear pore complexes embedded in the nuclear envelope. The nuclear lamina is associated with the inner face of the bilayer nuclear envelope whereas the outer face stays continuous with the endoplasmic reticulum.

Architecture of Nuclear Lamina

The nuclear lamina consists of two main components, lamins and nuclear lamin associated membrane proteins. The lamins are type V intermediate filaments which can be categorized as either A-type (lamin A, C) or B-type(lamin B1, B2) according to homology in sequence, biochemical properties and cellular localization during the cell cycle. Type V intermediate filaments differ from cytoplasmic intermediate filaments in the way that they have an extended rod domain (42 amino acid longer), that they all carry a nuclear localization signal (NLS) at their C-terminus and that they display typical tertiary structures. Lamin polypeptides have an almost complete α -helical conformation with multiple α -helical domains separated by non- α -helical linkers that are highly conserved in length and amino acid sequence. Both the C-terminus and the N-terminus are non α -helical, with the C-terminus displaying a globular structure. Their molecular weight ranges from 60 to 80 kilodaltons (kDa). In the amino acid sequence of nuclear lamins, there are also two phosphoacceptor sites present, flanking the central rod domain. A phosphorylation event at the onset of mitosis leads to a conformational change which causes the disassembly of

the nuclear lamina. In the vertebrate genome, lamins are encoded by three genes. By alternative splicing, at least seven different polypeptides (splice variants) are obtained, some of which are specific for germ cells and play an important role in the chromatin reorganisation during meiosis. Not all organisms have the same number of lamin encoding genes; *Drosophila melanogaster* for example has only 2 genes, whereas *Caenorhabditis elegans* has only one. The presence of lamin polypeptides is an exclusive property of Metazoan organisms. Plants or single-cell Eukaryotic organisms such as *Saccharomyces cerevisiae* lack lamins. The nuclear lamin-associated membrane proteins are either integral or peripheral membrane proteins. The most important are lamin associated polypeptide 1 and 2 (LAP1, LAP2), emerin, lamin B-receptor (LBR), otefin and MAN1. Due to their positioning within or their association with the inner membrane, they mediate the attachment of the nuclear lamina to the nuclear envelope.^[7]

Function of nuclear lamin

The nuclear lamina is assembled by interactions of two lamin polypeptides in which the α -helical regions are wound around each other to form a two stranded α -helical coiled-coil structure, followed by a head-to-tail association of the multiple dimers. The linearly elongated polymer is extended laterally by a side-by-side association of polymers, resulting in a 2D structure underlying the nuclear envelope. Next to providing mechanical support to the nucleus, the nuclear lamina plays an essential role in chromatin organization, cell cycle regulation, DNA replication, cell differentiation and apoptosis.

Chromatin organization The non-random organization of the genome strongly suggests that the nuclear lamina plays a role in chromatin organization. Indeed, it has been shown that lamin polypeptides have an affinity for binding chromatin through their α -helical (rod like) domains at specific DNA sequences called matrix attachment regions (MAR). A MAR has a length of approximately 300–1000 bp and has a high A/T content. Lamin A and B can also

bind core histones through a sequence element in their tail domain.^[8]

Cell cycle regulation At the onset of mitosis, (prophase, prometaphase) the cellular machinery is engaged in the disassembly of various cellular components including structures such as the nuclear envelope, the nuclear lamina and the nuclear pore complexes. This nuclear breakdown is necessary to allow the mitotic spindle to interact with the (condensed) chromosomes and to bind them at their kinetochores. These different disassembly events are initiated by the cyclin B/Cdk1 protein kinase complex (MPF). Once this complex is activated, the cell is forced into mitosis, by the subsequent activation and regulation of other protein kinases or by direct phosphorylation of structural proteins involved in this cellular reorganisation. After phosphorylation by cyclin B/Cdk1, the nuclear lamina depolymerises and B-type lamins stay associated with the fragments of the nuclear envelope whereas A-type lamins remain completely soluble throughout the remaining of the mitotic phase. The importance of the nuclear lamina breakdown at this stage is underlined by experiments where inhibition of the disassembly event leads to a complete cell cycle arrest. At the end of mitosis, (anaphase, telophase) there is a nuclear reassembly which is highly regulated in time, starting with the association of 'skeletal' proteins on the surface of the still partially condensed chromosomes, followed by nuclear envelope assembly. Novel nuclear pore complexes are formed through which nuclear lamins are actively imported by use of their NLS. This typical hierarchy raises the question whether the nuclear lamina at this stage has a stabilizing role or some regulative function, for it is clear that it plays no essential part in the nuclear membrane assembly around chromatin.

Embryonic development and cell differentiation The presence of lamins in embryonic development is readily observed in various model organisms such as *Xenopus laevis*, the chick and mammals. In *Xenopus laevis*, five different types were identified which are present in different expression patterns during the different stages

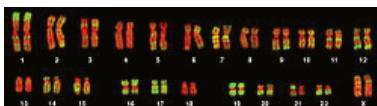
of the embryonic development. The major types are LI and LII, which are considered homologs of lamin B1 and B2. LA are considered homologous to lamin A and LIII as a B-type lamin. A fourth type exists and is germ cell specific. In the early embryonic stages of the chick, the only lamins present are B-type lamins. In further stages, the expression pattern of lamin B1 decreases and there is a gradual increase in the expression of lamin A. Mammalian development seems to progress in a similar way. In the latter case as well it is the B-type lamins that are expressed in the early stages. Lamin B1 reaches the highest expression level, whereas the expression of B2 is relatively constant in the early stages and starts to increase after cell differentiation. With the development of the different kinds of tissue in a relatively advanced developmental stage, there is an increase in the levels of lamin A and lamin C. These findings would indicate that in its most basic form, a functional nuclear lamina requires only B-type lamins.^[8]

DNA replication Various experiments show that the nuclear lamina plays a part in the elongation phase of DNA replication. It has been suggested that lamins provide a scaffold, essential for the assembly of the elongation complexes, or that it provides an initiation point for the assembly of this nuclear scaffold. Not only nuclear lamina associated lamins are present during replication, but free lamin polypeptides are present as well and seem to have some regulative part in the replication process.^[8]

Apoptosis Apoptosis, basically to be considered as cellular suicide is of the highest importance in homeostasis of tissue and in defending the organism against invasive entry of viruses or other pathogens. Apoptosis is a highly regulated process in which the nuclear lamina is disassembled in an early stage. In contrast to the phosphorylation-induced disassembly during mitosis, the nuclear lamina is degraded by proteolytic cleavage, and both the lamins and the nuclear lamin-associated membrane proteins are targeted. This proteolytic activity is performed by members of the caspase-protein family who cleave the lamins after aspartic acid (Asp) residues.^{[7][8]}

Chromosomes

In a series of experiments beginning in the mid-1880s, Theodor Boveri gave the definitive demonstration that chromosomes are the vectors of heredity. His two principles were the continuity of chromosomes and the individuality of chromosomes. It is the second of these principles that was so original. Wilhelm Roux suggested that each chromosome carries a different genetic load. Boveri was able to test and confirm this hypothesis. Aided by the rediscovery at the start of the 1900s of Gregor Mendel's earlier work, Boveri was able to point out the connection between the rules of inheritance and the behaviour of the chromosomes. 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 in Development and Heredity*, Wilson linked together the independent work of Boveri and Sutton (both around 1902) by naming the chromosome theory of inheritance the "Sutton-Boveri Theory" (the names are sometimes reversed). Ernst 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. The cell nucleus contains the majority of the cell's genetic material, in the form of multiple linear DNA molecules organized into structures called chromosomes. 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



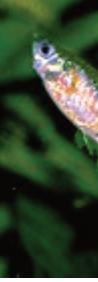
Karyogram from a human female lymphocyte probed for the Alu sequence using FISH.

package the DNA and control its functions. The word chromosome^[9] comes from the Greek χρῶμα (chroma, colour) and σῶμα (soma, body) due to their property of being very strongly stained by particular dyes.^[10]

Telomere A telomere is a region of repetitive DNA at the end of a chromosome, which protects the end of the chromosome from deterioration. Its name is derived from the Greek nouns telos “end” and meros “part”. The telomere regions deter the degradation of genes near the ends of chromosomes by allowing for the shortening of chromosome ends, which necessarily occurs during chromosome replication.

The total number of chromosomes (including sex chromosomes) present in a cell nucleus of different organisms are described below in the tables.^[11]

Chromosome numbers in some plants^[11]

Plant Species	# ^[11]	Species
Arabidopsis thaliana (diploid) ^[12]		Common fruit fly
	10	 Male (left) and
Rye (diploid) ^[13]	14	 Guppy (Poecilia

Maize (diploid or palaeotetraploid)^[14]



20

Stalks, ears, and silk

Earthworm (O



Ocyurus olens tr

Einkorn wheat (diploid)^[15]

14

Durum wheat (tetraploid)^[15]

28

Bread wheat (hexaploid)^[15]

Domestic cat



42

Ears of compact wheat

Laboratory m



BALB/c mice

Cultivated tobacco (tetraploid)^[16]

48

Adder's Tongue Fern (diploid)^[17]



approx. 1,200

Rabbit (Orycto)

Hares^{[26][27]}

Gorillas, Chim

Elephants^[29]

Donkey

Dog^[30]

Goldfish^[32]

Chromosome numbers in other organisms^[11]

Species	Large Chromosomes	Intermediate Chromosomes	Microchrom
<i>Trypanosoma brucei</i>	11	6	~100
Domestic Pigeon (<i>Columba livia domestica</i>) ^[34]	18	–	59–63
Chicken ^[35]	8	2 sex chromosomes	60

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.^[11]

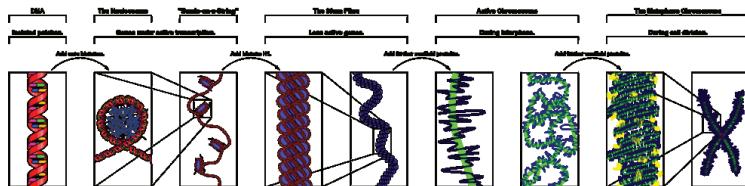


Fig. 2: The major structures in DNA compaction; DNA, the nucleosome, the 10nm “beads-on-a-string” fibre, the 30nm fibre and the metaphase chromosome.

Nucleolus

The nucleolus is a discrete densely stained structure found in the nucleus. It is non-membrane bound structure, and is sometimes called a suborganelle. **It forms around tandem repeats of rDNA**, DNA coding for ribosomal RNA (rRNA). These regions are called nucleolar organizer regions (NOR). The main roles of the nucleolus are to synthesize rRNA and assemble ribosomes. The structural cohesion of the nucleolus depends on its activity, as ribosomal assembly in the nucleolus results in the transient association of nucleolar components, facilitating further ribosomal assembly, and hence further association. This model is supported by observations that inactivation of rDNA results in intermingling of nucleolar structures.^[36]

Splicing speckles or SC35 speckles

Sometimes referred to as interchromatin granule clusters (IGCs) or as splicing-factor compartments, speckles are rich in splicing snRNPs and other splicing proteins necessary for pre-mRNA processing. Because of a cell's changing requirements, the composition and location of these bodies changes according to

mRNA transcription and regulation via phosphorylation of specific proteins.^[36]

Cajal bodies or coiled bodies

A nucleus typically contains between 1 and 10 compact structures called Cajal bodies or coiled bodies (CB), whose diameter measures between 0.2 µm and 2.0 µm depending on the cell type and species. When seen under an electron microscope, they resemble balls of tangled thread and are dense foci of distribution for the protein coilin. CBs are involved in a number of different roles relating to RNA processing, specifically small nucleolar RNA (snoRNA) and small nuclear RNA (snRNA) maturation, and histone mRNA modification.^[36]

Nuclear import and export

Nuclear localization signal (NLS)

A nuclear localization signal or sequence (NLS) is an amino acid sequence which acts like a ‘tag’ on the exposed surface of a protein. This sequence is used to target the protein to the cell nucleus through the Nuclear Pore Complex and to direct a newly synthesized protein into the nucleus via its recognition by cytosolic nuclear transport receptors. Typically, this signal consists of one or more short sequences of positively charged lysines or arginines. Different nuclear localized proteins may share the same NLS. An NLS has the opposite function of a nuclear export signal, which targets proteins out of the nucleus.^[37]

A. Classical NLSs

Classical Nuclear localization signals can be further classified as either monopartite or bipartite. **The first NLS to be discovered was the sequence PKKKRKV in the SV40 Large T-antigen (a monopartite NLS).** The NLS of nucleoplasmmin, KR[PAATKKAGQA]KKKK, is the prototype of the ubiquitous bipartite signal: two clusters of basic amino acids, separated by a spacer of about 10 amino acids. Both signals are recognized by importin α . Importin α contains a bipartite NLS itself, which is specifically recognized by importin β . The latter can be considered the actual import mediator. Chelsky et al. proposed the consensus sequence K-K/R-X-K/R for monopartite NLSs. A Chelsky sequence may, therefore, be part of the downstream basic cluster of a bipartite NLS. Makkerh et al. carried out comparative mutagenesis on the nuclear localisation signals of SV40 T-Antigen (monopartite), C-myc (monopartite) and nucleoplasmmin (bipartite), and showed amino acid features common to all three. Notably the role of neutral and acidic amino acids was shown for the first time in contributing to the efficiency of the NLS.^[37]

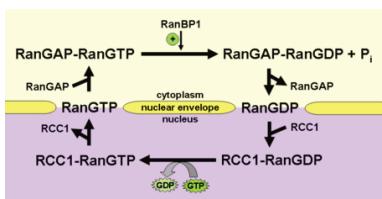
B. Non-classical NLSs

There are many other types of NLS, such as the acidic M9 domain of hnRNP A1, the sequence KIPIK in yeast transcription repressor Mata α 2, and the complex signals of U snRNPs. Most of these NLSs appear to be recognized directly by specific receptors of the importin β family without the intervention of an importin α -like protein. A signal that appears to be specific for the massively produced and transported ribosomal proteins, seems to come with a specialized set of importin β -like nuclear import receptors. Recently a class of NLSs known as PY-NLSs has been proposed, originally by Lee et al. This PY-NLS motif, so named because of the proline-tyrosine amino acid pairing in it, allows the protein to bind to Importin β 2 (also known as transportin or karyopherin β 2), which then translocates the cargo protein into the nucleus. The structural basis for the binding of the PY-NLS contained in Importin β 2 has been determined and an inhibitor of import designed.^[37]

Nuclear export signal (NES)

A nuclear export signal (NES) is a short amino acid sequence of 4 hydrophobic residues in a protein that targets it for export from the cell nucleus to the cytoplasm through the nuclear pore complex. It has the opposite effect of a nuclear localization signal, which targets a protein located in the cytoplasm for import to the nucleus. The NES is recognized and bound by exportins. In silico analysis of known NESs found the most common spacing of the hydrophobic residues to be LxxxLxxLxL, where “L” is a hydrophobic residue (often leucine) and “x” is any other amino acid; the spacing of these hydrophobic residues may be explained by examination of known structures that contain an NES, as the critical residues usually lie in the same face of adjacent secondary structures within a protein, which allows them to interact with the exportin[1]. Ribonucleic acid (RNA) are composed of nucleotides, and thus, lack the nuclear export signal to move out of the nucleus. As a result, most forms of RNA will bind to a protein molecule to form a ribonucleoprotein complex to be exported from the nucleus. Nuclear export first begins with the binding of Ran-GTP (a G-protein) to exportin. This causes a shape change in exportin, increasing its affinity for the export cargo. Once the cargo is bound, the Ran-exportin-cargo complex moves out of the nucleus through the nuclear pore. GTPase activating proteins (GAPs) then hydrolyze the Ran-GTP to Ran-GDP, and this causes a shape change and subsequent exportin release. Once no longer bound to Ran, the exportin molecule loses affinity for the nuclear cargo as well, and the complex falls apart. Exportin and Ran-GDP are recycled to the nucleus separately, and guanine exchange factor (GEF) in the nucleus switches the GDP for GTP on Ran.^[37]

Role of Ran GTPase in nuclear transport during interphase

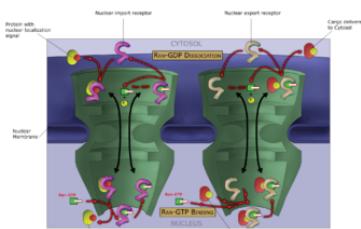


Schematic representation of the Ran cycle

cell in two nucleotide-bound forms: GDP-bound and GTP-bound. RanGDP is converted into RanGTP through the action of RCC1 (regulator of chromosome condensation 1), the nucleotide exchange factor for Ran. RCC1 is also known as RanGEF (Ran Guanine nucleotide Exchange Factor). Ran's intrinsic GTPase-activity is activated through interaction with Ran GTPase activating protein (RanGAP), facilitated by complex formation with Ran-binding protein (RanBP). GTPase-activation leads to the conversion of RanGTP to RanGDP, thus closing the Ran cycle.

Ran can diffuse freely within the cell, but because RCC1 and RanGAP are located in different places in the cell, the concentration of RanGTP and RanGDP differs locally as well, creating concentration gradients that act as signals for other cellular processes. RCC1 is bound to chromatin and therefore located inside the nucleus. RanGAP is cytoplasmic in yeast and bound to the nuclear envelope in plants and animals. In mammalian cells, it is SUMO modified and attached to the cytoplasmic side of the nuclear pore complex via interaction with the nucleoporin RanBP2 (Nup358). This difference in location of the accessory proteins in the Ran cycle

Ran (RAs-related Nuclear protein) is a small 25Kda protein that is involved in transport into and out of the cell nucleus during interphase and also involved in mitosis. It is a member of the Ras superfamily. Ran exists in the



Ran cycle involvement in nucleocytoplastic transport at the nuclear pore

leads to a high RanGTP to RanGDP ratio inside the nucleus and an inversely low RanGTP to RanGDP ratio outside the nucleus. In addition to a gradient of the nucleotide bound state of Ran, there is a gradient of the protein itself, with a higher concentration of Ran in the nucleus than in the cytoplasm. Cytoplasmic RanGDP is imported into the nucleus by the small protein NTF2 (Nuclear Transport Factor 2), where RCC1 can then catalyze exchange of GTP for GDP on Ran.

Ran is involved in the transport of proteins across the nuclear envelope by interacting with karyopherins and changing their ability to bind or release cargo molecules. Cargo proteins containing a nuclear localization signal (NLS) are bound by importins and transported into the nucleus. Inside the nucleus, RanGTP binds to importin and releases the import cargo. Cargo that needs to get out of the nucleus into the cytoplasm binds to exportin in a ternary complex with RanGTP. Upon hydrolysis of RanGTP to RanGDP outside the nucleus, the complex dissociates and export cargo is released.^[38]

Importin

Importin also play important role in nuclear transport. Importin is a type of protein that moves other protein molecules into the nucleus by binding to a specific recognition sequence, called the nuclear localization signal (NLS). Importin is classified as a karyopherin. Importin has two subunits, importin α and importin β . Of these, importin α binds to the NLS of the protein to be imported to the nucleus whereas importin β helps in the docking of the importin heterodimer-bound protein to the nuclear pore complex. The NLS-Importin α -Importin β trimer dissociates after binding to Ran GTP inside the nucleus^[39]

Function of nucleus

The main function of the cell nucleus is to control gene expression and mediate the replication of DNA during the cell cycle. The nucleus provides a site for genetic transcription that is segregated from the location of translation in the cytoplasm,

allowing levels of gene regulation that are not available to prokaryotes.

Cell compartmentalization

The nuclear envelope allows the nucleus to control its contents, and separate them from the rest of the cytoplasm where necessary. This is important for controlling processes on either side of the nuclear membrane. In some cases where a cytoplasmic process needs to be restricted, a key participant is removed to the nucleus, where it interacts with transcription factors to downregulate the production of certain enzymes in the pathway. This regulatory mechanism occurs in the case of glycolysis, a cellular pathway for breaking down glucose to produce energy. Hexokinase is an enzyme responsible for the first step of glycolysis, forming glucose-6-phosphate from glucose. At high concentrations of fructose-6-phosphate, a molecule made later from glucose-6-phosphate, a regulator protein removes hexokinase to the nucleus, where it forms a transcriptional repressor complex with nuclear proteins to reduce the expression of genes involved in glycolysis. In order to control which genes are being transcribed, the cell separates some transcription factor proteins responsible for regulating gene expression from physical access to the DNA until they are activated by other signaling pathways. This prevents even low levels of inappropriate gene expression. For example, in the case of NF- κ B-controlled genes, which are involved in most inflammatory responses, transcription is induced in response to a signal pathway such as that initiated by the signaling molecule TNF- α , binds to a cell membrane receptor, resulting in the recruitment of signalling proteins, and eventually activating the transcription factor NF- κ B. A nuclear localisation signal on the NF- κ B protein allows it to be transported through the nuclear pore and into the nucleus, where it stimulates the transcription of the target genes. The compartmentalization allows the cell to prevent translation of unspliced mRNA. Eukaryotic mRNA contains introns that must be removed before being translated to produce functional proteins. The splicing is done inside the nucleus before the mRNA can be

accessed by ribosomes for translation. Without the nucleus, ribosomes would translate newly transcribed (unprocessed) mRNA, resulting in misformed and nonfunctional proteins.^[2]

Gene expression

Gene expression first involves transcription, in which DNA is used as a template to produce RNA. In the case of genes encoding proteins, that RNA produced from this process is messenger RNA (mRNA), which then needs to be translated by ribosomes to form a protein. As ribosomes are located outside the nucleus, mRNA produced needs to be exported. Since the nucleus is the site of transcription, it also contains a variety of proteins that either directly mediate transcription or are involved in regulating the process. These proteins include helicases, which unwind the double-stranded DNA molecule to facilitate access to it, RNA polymerases, which synthesize the growing RNA molecule, topoisomerases, which change the amount of supercoiling in DNA, helping it wind and unwind, as well as a large variety of transcription factors that regulate expression.^[2]

Processing of pre-mRNA

Newly synthesized mRNA molecules are known as primary transcripts or pre-mRNA. They must undergo post-transcriptional modification in the nucleus before being exported to the cytoplasm; mRNA that appears in the cytoplasm without these modifications is degraded rather than used for protein translation. The three main modifications are 5' capping, 3' polyadenylation, and RNA splicing. While in the nucleus, pre-mRNA is associated with a variety of proteins in complexes known as heterogeneous ribonucleoprotein particles (hnRNPs). Addition of the 5' cap occurs co-transcriptionally and is the first step in post-transcriptional modification. The 3' poly-adenine tail is only added after transcription is complete. RNA splicing, carried out by a complex called the spliceosome, is the process by which introns, or regions of DNA that do not code for protein, are removed from the pre-mRNA and the remaining exons connected to re-form a single continuous molecule. This process normally occurs after 5' capping.

and 3' polyadenylation but can begin before synthesis is complete in transcripts with many exons. Many pre-mRNAs, including those encoding antibodies, can be spliced in multiple ways to produce different mature mRNAs that encode different protein sequences. This process is known as alternative splicing, and allows production of a large variety of proteins from a limited amount of DNA.^[2]

Facts to be remembered

Subnuclear structure sizes

Structure name	Structure diameter	
Cajal bodies	0.2–2.0 µm	[40]
PIKA	5 µm	[41]
PML bodies	0.2–1.0 µm	[42]
Paraspeckles	0.2–1.0 µm	
Speckles	20–25 nm	[41]

Red blood cell of human doesn't contain nucleus.

Intestinal parasites in the genus *Giardia* which have **two nuclei** per cell.

Some species of protozoa and some fungi in mycorrhizae have polynucleated cells.

During mitosis the NPC appears to disassemble in stages. Peripheral nucleoporins such as the Nup 153 Nup 98 and Nup 214 disassociate from the NPC. The rest, which can be considered a scaffold proteins remain stable, as cylindrical ring complexes within the nuclear envelope. This disassembly of the NPC peripheral groups is largely thought to be phosphate driven, as several of these nucleoporins are phosphorylated during the stages of mitosis.

However, the enzyme involved in the phosphorlyation is unknown in vivo.

In metazoans (which undergo open mitosis) the NE degrades quickly after the loss of the peripheral Nups. The reason for this may be due to the change in the NPC's architecture. This change may make the NPC more permeable to enzymes involved in the degradation of the NE such as cytoplasmic tubulin, as well as allowing the entry of key mitotic regulator proteins.

The table given below lists the numbers of chromosomes in various plants, animals, protists, and other living organisms, given as the diploid number ($2n$)^[43]

Organism	Scientific name	Diploid number of chromosomes	Notes
African Wild Dog	<i>Lycaon pictus</i>	78 ^[44]	
Alfalfa	<i>Medicago sativa</i>	32 ^[45]	Cultivated alfalfa is tetraploid, with $2n=4x=32$. Wild relatives have $2n=16$. ^[45]
American Badger		32	
American Marten		38	
American Mink		30	
Aquatic Rat	<i>Anatomys leander</i>	92 ^[46]	Tied for highest number in mammals with <i>Ichthyomys pittieri</i> .
Arabidopsis thaliana		10	
Barley	<i>Hordeum vulgare</i>	14	
Bat-eared Fox	<i>Otocyon megalotis</i>	72 ^[44]	
Bean	<i>Phaseolus</i> sp.	22 ^[45]	All species in the genus have the same chromosome number, including <i>P. vulgaris</i> , <i>P. coccineus</i> , <i>P. acutifolius</i> , and <i>P. lunatus</i> . ^[45]
Beaver (American)	<i>Castor canadensis</i>	40	
Beaver (Eurasian)	<i>Castor fiber</i>	48	
Beech Marten		38	
Bengal Fox	<i>Vulpes bengalensis</i>	60	
Moonworts	<i>Botrychium</i>	90	

Organism	Scientific name	Diploid number of chromosomes	Notes
nagaho-no-nastu-no-hana-warabi	<i>Botrypus strictus</i>	88	<i>B. strictus</i> and <i>B. virginianus</i> have been shown to be paraphyletic in the genus <i>Botrypus</i>
Rattlesnake fern	<i>Botrypus virginianus</i>	184	
Cabbage	<i>Brassica oleracea</i>	18 ^[45]	Broccoli, cabbage, kale, kohlrabi, brussels sprouts, and cauliflower are all the same species and have the same chromosome number.
Carp		104	
Capuchin Monkey		54 ^[47]	
Cat	<i>Felis catus</i>	38	
Chicken	<i>Gallus gallus domesticus</i>	78	
Chimpanzee	<i>Pan troglodytes</i>	48 ^[48]	
Chinchilla	<i>Chinchilla lanigera</i>	64 ^[49]	
Coatimundi		38	
Cotton	<i>Gossypium hirsutum</i>	52 ^[45]	2n=4x; Cultivated upland cotton is derived from an allotetraploid
Cow	<i>Bos primigenius</i>	60	
Coyote	<i>Canis latrans</i>	78 ^[44]	
Deer Mouse		48	
Dhole		78	

Organism	Scientific name	Diploid number of chromosomes	Notes
Dingo	<i>Canis lupus dingo</i>	78 ^[44]	
Dog	<i>Canis lupus familiaris</i>	78 ^[50]	76 autosomal and 2 sexual. ^[51]
Dolphin	<i>Delphinidae</i> <i>Delphis</i>	44	
Donkey		62	
Dove		78 ^[52]	Based on African collared dove
Fruit fly	<i>Drosophila melanogaster</i>	8 ^[53]	6 autosomal, and 2 sexual
Duck-billed Platypus		52	
Earthworm	<i>Lumbricus terrestris</i>	36	
Echidna		63/64	63 (XXY, male) and 64 (XXXX, female)
Elephant		56	
Elk (Wapiti)	<i>Cervus canadensis</i>	68	
Eurasian Badger		44	
European honey bee	<i>Apis mellifera</i>	32	32 for females, males are haploid and thus have 16.
European Mink		38	
European Polecat		40	
Fennec Fox	<i>Vulpes zerda</i>	64 ^[44]	
Ferret		40	
Field Horsetail		216	
Fisher (animal)		38	a type of marten
Fossa		42	

Organism	Scientific name	Diploid number of chromosomes	Notes
Giraffe	<i>Giraffa camelopardalis</i>	62	
Goat		60	
Golden Jackal	<i>Canis aureus</i>	78 ^[44]	
Gorilla		48	
Gray Fox	<i>Urocyon cinereoargenteus</i>	66 ^[44]	
Gypsy moth		62	
Hawkweed		8	
Hare ^{[54][55]}		48	
Hedgehog Genus Atelerix (African hedgehogs)		90	
Hedgehog Genus Erinaceus (Woodland hedgehogs)		88	
fern-like plant	<i>Helminthostachys zeylanica</i>	94	
Horse	<i>Equus ferus caballus</i>	64	
Human	<i>Homo sapiens</i>	46 ^[56]	44 autosomal and 2 sex
Hyena		40	
Crab-eating rat (se미수생 rodent)	<i>Ichthyomys pittieri</i>	92 ^[46]	highest for a mammal

Organism	Scientific name	Diploid number of chromosomes	Notes
Jack jumper ant	<i>Myrmecia pilosula</i>	2 ^[57]	2 for females, males are haploid and thus have 1; smallest number possible. Other ant species have more chromosomes. ^[57]
Kangaroo		12	
Kit Fox		50	
Lion	<i>Panthera leo</i>	38	
Long-nosed Cusimanse (a type of mongoose)		36	
Maize	<i>Zea mays</i>	20 ^[45]	
Maned Wolf	<i>Chrysocyon brachyurus</i>	76	
Mango	<i>Mangifera indica</i>	40 ^[45]	
Meerkat		36	
Mosquito	<i>Aedes aegypti</i>	6 ^[58]	The 2n=6 chromosome number is conserved in the entire family Culicidae, except in <i>Chagasia bathana</i> which has 2n=8. ^[58]
Mouse	<i>Mus musculus</i>	40	
Mule		63	semi-infertile

Organism	Scientific name	Diploid number of chromosomes	Notes
Oats	<i>Avena sativa</i>	42 ^[45]	This is a hexaploid with $2n=6x=42$. Diploid and tetraploid cultivated species also exist. ^[45]
Adders-tongue	<i>Ophioglossum reticulatum</i>	1200 or 1260	This fern has the highest known chromosome number.
Orangutan		48	
Oriental Small-clawed Otter		38	
Pea	<i>Pisum sativum</i>	14	
Pig		38	
Pigeon		80	
Pine Marten		38	
Pineapple	<i>Ananas comosus</i>	50 ^[45]	
Potato	<i>Solanum tuberosum</i>	48	This is a tetraploid; wild relatives mostly have $2n=24$. ^[45]
Porcupine	<i>Erethizon dorsatum</i>	34 ^[49]	
Rabbit		44	
Raccoon (<i>Procyon lotor</i>)		38 ^[59]	
Raccoon Dog	<i>Nyctereutes viverrinus</i>	42	some sources say sub-species differ with 38, 54 and even 56 chromosomes
Raccoon Dog	<i>Nyctereutes procyonoides</i>	56	

Organism	Scientific name	Diploid number of chromosomes	Notes
Radish	<i>Raphanus sativus</i>	18	
Rat		42	
Red Deer	<i>Cervus elaphus</i>	68	
Red Fox	<i>Vulpes vulpes</i>	34 ^[44]	Plus 3-5 microsomes.
Red Panda		36	
Reeves's Muntjac	<i>Muntiacus reevesi</i>	46	
Rice	<i>Oryza sativa</i>	24 ^[45]	
Rhesus Monkey		48	
Rye	<i>Secale cereale</i>	14 ^[45]	
Sable		38	
Sable Antelope		46	
Grape ferns	<i>Sceptridium</i>	90	
Sea Otter		38	
Sheep		54	
Shrimp	<i>Penaeus semisulcatus</i>	86-92 ^[60]	
Slime Mold	<i>Dictyostelium discoideum</i>	12 ^[61]	
Snail		24	
Spotted Skunk		64	
Starfish		36	
Striped skunk		50	
Swamp Wallaby	<i>Wallabia bicolor</i>	10/11	10 for male, 11 for female

Organism	Scientific name	Diploid number of chromosomes	Notes
Tanuki/Raccoon Dog	<i>Nyctereutes procyonoides albus</i>	38	
Tiger	<i>Panthera tigris</i>	38	
Tibetan fox		36	
Tobacco	<i>Nicotiana tabacum</i>	48	Cultivated species is a tetraploid. ^[45]
Turkey		82	
Virginia Opossum	<i>Didelphis virginiana</i>	22 ^[62]	
Wheat	<i>Triticum aestivum</i>	42 ^[45]	This is a hexaploid with $2n=6x=42$. Durum wheat is <i>Triticum turgidum</i> var. <i>durum</i> , and is a tetraploid with $2n=4x=28$. ^[45]
White-tailed deer	<i>Odocoileus virginianus</i>	70	
Wolf		78	
Woolly Mammoth		58	extinct; tissue from a frozen carcass
Wolverine		42	
Yellow Mongoose		36	
Yeast		32	
Bittersweet nightshade	<i>Solanum dulcamara</i>	24 ^{[63][64]}	
Husk Tomato	<i>Physalis pubescens</i>	24 ^[65]	

Question time

1. What do you know about nuclear membrane?
2. What is the function of nucleolus?
3. What is a chromosome?
4. Design an experiment which indicate that a particular protein is going into nucleus?
5. How will you determine that a particular protein X is localizing in to the nucleus not in cytoplasm?
6. What are importins?
7. What due understand with NLS?
8. What is the difference between active and passive transport?

References

1. ↑ Cell Nucleus
2. ↑ Jump up to: **a b c d** Cell nucleus
3. ↑ Chi YH, Chen ZJ, Jeang KT (2009). "The nuclear envelopathies and human diseases". *J. Biomed. Sci.* **16**: 96. doi:10.1186/1423-0127-16-96. PMID 19849840. PMC 2770040. <http://www.jbiomedsci.com/content/16/96>.
4. ↑ Nuclear envelope
5. ↑ Nuclear envelope
6. ↑ Jump up to: **a b** Nuclear pore
7. ↑ Jump up to: **a b** Nuclear lamina
8. ↑ Jump up to: **a b c d** http://en.wikipedia.org/wiki/Nuclear_lamina
9. ↑ Bolzer et al., (2005) Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biol* 3(5)
10. ↑ Chromosome
11. ↑ Jump up to: **a b c d e f** <http://en.wikipedia.org/wiki/>

Chromosome

12. ↑ Armstrong SJ, Jones GH (January 2003). “Meiotic cytology and chromosome behaviour in wild-type *Arabidopsis thaliana*”. *J. Exp. Bot.* **54** (380): 1–10. doi:10.1093/jxb/54.380.1. PMID 12456750. <http://jexbot.oxfordjournals.org/cgi/pmidlookup?view=long&pmid=12456750>.
13. ↑ Gill BS, Kimber G (April 1974). “The Giemsa C-banded karyotype of rye”. *Proc. Natl. Acad. Sci. U.S.A.* **71** (4): 1247–9. doi:10.1073/pnas.71.4.1247. PMID 4133848.
14. ↑ Kato A, Lamb JC, Birchler JA (September 2004). “Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize”. *Proc. Natl. Acad. Sci. U.S.A.* **101** (37): 13554–9. doi:10.1073/pnas.0403659101. PMID 15342909. PMC 518793. <http://www.pnas.org/cgi/pmidlookup?view=long&pmid=15342909>.
15. ↑ Jump up to: **a b c** Dubcovsky J, Luo MC, Zhong GY, et al. (1996). “Genetic map of diploid wheat, *Triticum monococcum* L., and its comparison with maps of *Hordeum vulgare* L”. *Genetics* **143** (2): 983–99. PMID 8725244.
16. ↑ Kenton A, Parokonny AS, Gleba YY, Bennett MD (August 1993). “Characterization of the *Nicotiana tabacum* L. genome by molecular cytogenetics”. *Mol. Gen. Genet.* **240** (2): 159–69. doi:10.1007/BF00277053. PMID 8355650.
17. ↑ Leitch IJ, Soltis DE, Soltis PS, Bennett MD (2005). “Evolution of DNA amounts across land plants (embryophyta)”. *Ann. Bot.* **95** (1): 207–17. doi:10.1093/aob/mci014. PMID 15596468. <http://aob.oxfordjournals.org/cgi/content/full/95/1/207>.
18. ↑ Umeko Semba, Yasuko Umeda, Yoko Shibuya, Hiroaki Okabe, Sumio Tanase and Tetsuro Yamamoto (2004). “Primary structures of guinea pig high- and low-molecular-weight kininogens”. *International Immunopharmacology* **4** (10-11): 1391–1400. doi:10.1016/j.intimp.2004.06.003. PMID 15313436. <http://www.sciencedirect.com/science/article/B6W7N-4CX6PRC-1/2/97487fdf611a04e3c88690da9e6d853b>.
19. ↑ “The Genetics of the Popular Aquarium Pet – Guppy Fish”.

- <http://fancyguppy.webs.com/genetics.htm>. Retrieved 2009-12-06.
- 20. ↑ Vitturi R, Libertini A, Sineo L, et al. (2005). "Cytogenetics of the land snails *Cantareus aspersus* and *C. mazzullii* (Mollusca: Gastropoda: Pulmonata)". *Micron* **36** (4): 351–7. doi:10.1016/j.micron.2004.12.010. PMID 15857774.
 - 21. ↑ Vitturi R, Colombo MS, Pirrone AM, Mandrioli M (2002). "rDNA (18S-28S and 5S) colocalization and linkage between ribosomal genes and (TTAGGG)(n) telomeric sequence in the earthworm, *Octodrilus complanatus* (Annelida: Oligochaeta: Lumbricidae), revealed by single- and double-color FISH". *J. Hered.* **93** (4): 279–82. doi:10.1093/jhered/93.4.279. PMID 12407215. <http://jhered.oxfordjournals.org/cgi/content/full/93/4/279>.
 - 22. ↑ Nie W, Wang J, O'Brien PC, et al. (2002). "The genome phylogeny of domestic cat, red panda and five mustelid species revealed by comparative chromosome painting and G-banding". *Chromosome Res.* **10** (3): 209–22. doi:10.1023/A:1015292005631. PMID 12067210.
 - 23. ↑ Jump up to: **a b** Romanenko, Svetlana A; Polina L Perelman, Natalya A Serdukova, Vladimir A Trifonov, Larisa S Biltueva, Jinhuan Wang, Tangliang Li, Wenhui Nie, Patricia C. M. O'Brien, Vitaly T. Volobouev, Roscoe Stanyon, Malcolm A. Ferguson-Smith, Fengtang Yang, Alexander S. Graphodatsky (2006-12). "Reciprocal chromosome painting between three laboratory rodent species". *Mammalian Genome: Official Journal of the International Mammalian Genome Society* **17** (12): 1183–1192. doi:10.1007/s00335-006-0081-z. ISSN 0938-8990. PMID 17143584. <http://www.ncbi.nlm.nih.gov/pubmed/17143584>. Retrieved 2009-10-14.
 - 24. ↑ Jump up to: **a b** A Comparison of the chromosomes of the rat and mouse with reference to the question of chromosome homology in mammals
 - 25. ↑ Hayes, H.; C. Rogel-Gaillard, C. Zijlstra, N.A. de Haan, C. Urien, N. Bourgeaux, M. Bertaud, A.A. Bosma (2002).

- “Establishment
of an R-banded rabbit karyotype nomenclature by FISH
localization of 23 chromosome-specific genes on both G- and
R-banded chromosomes”. *Cytogenetic and Genome Research* **98**
(2-3): 199–205. doi:10.1159/000069807. ISSN 1424-859X.
PMID 12698004. <http://content.karger.com/produktedb/produkte.asp?typ=fulltext&file=CGR98199>. Retrieved
2009-10-14.
26. ↑ T.J. Robinson, F. Yang, W.R. Harrison (2002). “Chromosome painting refines the history of genome evolution in hares and rabbits (order Lagomorpha)”. *Cytogenetic and Genetic Research* **96** (1-4): 223–227. doi:10.1159/000063034. PMID 12438803.
<http://content.karger.com/ProdukteDB/produkte.asp?Aktion>ShowAbstract&ArtikelNr=63034&Ausgabe=228416&ProduktNr=224037>.
27. ↑ “Rabbits, Hares and Pikas. Status Survey and Conservation Action Plan”. pp. 61–94.
<http://wildlife1.wildlifeinformation.org/s/00Ref/BooksContents/b605.htm>.
28. ↑ ^{Jump up to: a b} De Grouchy J (1987). “Chromosome phylogenies of man, great apes, and Old World monkeys”. *Genetica* **73** (1-2): 37–52. PMID 3333352.
29. ↑ Houck ML, Kumamoto AT, Gallagher DS, Benirschke K (2001). “Comparative cytogenetics of the African elephant (*Loxodonta africana*) and Asiatic elephant (*Elephas maximus*)”. *Cytogenet. Cell Genet.* **93** (3-4): 249–52. doi:10.1159/000056992.
PMID 11528120.
30. ↑ Wayne RK, Ostrander EA (1999). “Origin, genetic diversity, and genome structure of the domestic dog”. *Bioessays* **21** (3): 247–57. doi:10.1002/(SICI)1521-1878(199903)21:3 (inactive 2009-03-11). PMID 10333734.
31. ↑ Burt DW (2002). “Origin and evolution of avian microchromosomes”. *Cytogenet. Genome Res.* **96** (1-4): 97–112.
doi:10.1159/000063018. PMID 12438785.
32. ↑ Ciudad J, Cid E, Velasco A, Lara JM, Aijón J, Orfao A (2002).

- “Flow cytometry measurement of the DNA contents of G0/G1 diploid cells from three different teleost fish species”. *Cytometry* **48** (1): 20–5. doi:10.1002/cyto.10100. PMID 12116377.
33. ↑ Yasukochi Y, Ashakumary LA, Baba K, Yoshido A, Sahara K (2006). “A second-generation integrated map of the silkworm reveals synteny and conserved gene order between lepidopteran insects”. *Genetics* **173** (3): 1319–28. doi:10.1534/genetics.106.055541. PMID 16547103.
34. ↑ Itoh, Masahiro; Tatsuro Ikeuchi, Hachiro Shimba, Michiko Mori, Motomichi Sasaki, Sajiro Makino (1969). “A COMPARATIVE KARYOTYPE STUDY IN FOURTEEN SPECIES OF BIRDS”. *The Japanese journal of genetics* **44** (3): 163–170. doi:10.1266/jgg.44.163. ISSN 1880-5787.
http://www.journalarchive.jst.go.jp/english/jnlabstract_en.php?cdjournal=ggs1921&cdvol=44&noissue=3&startpage=163. Retrieved 2009-10-14.
35. ↑ Smith J, Burt DW (1998). “Parameters of the chicken genome (*Gallus gallus*)”. *Anim. Genet.* **29** (4): 290–4. doi:10.1046/j.1365-2052.1998.00334.x. PMID 9745667.
36. ↑ Jump up to: **a b c** http://en.wikipedia.org/wiki/Cell_nucleus
37. ↑ Jump up to: **a b c d** Nuclear localization signal
38. ↑ Ran (biology)
39. ↑ <http://en.wikipedia.org/wiki/Importin>
40. ↑ Cioce M, Lamond A. “Cajal bodies: a long history of discovery”. *Annu Rev Cell Dev Biol* **21**: 105–131. doi:10.1146/annurev.cellbio.20.010403.103738. PMID 16212489.
41. ↑ Jump up to: **a b** Pollard, Thomas D.; William C. Earnshaw (2004). *Cell Biology*. Philadelphia: Saunders. ISBN 0-7216-3360-9.
42. ↑ Dundr, Miroslav; Tom Misteli (2001). “Functional architecture in the cell nucleus”. *Biochem. J.* (356): 297–310. PMID 11368755.
43. ↑ List of organisms by chromosome count
44. ↑ Jump up to: **a b c d e f g h** Sillero-Zubiri, Claudio; Hoffmann, Michael J.; Dave Mech (2004). *Canids: Foxes, Wolves, Jackals and Dogs: Status Survey and Conservation Action Plan*. World Conservation Union. ISBN 2-8317-0786-2.

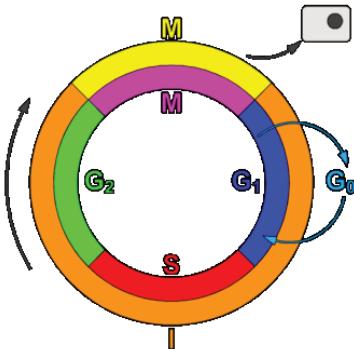
45. ↑ Jump up to: **a b c d e f g h i j k l m n o p q** Simmonds, NW (ed.) (1976). *Evolution of crop plants*. New York: Longman. ISBN 0-582-44496-9. Template:Page needed
46. ↑ Jump up to: **a b** Schmid, M.; Fernández-Badillo, A.; Feichtinger, W.; Steinlein, C.; Roman, J.I. (1988). "On the highest chromosome number in mammals". *Cytogenetics and Genome Research* **49** (4): 305–8. doi:10.1159/000132683. PMID 3073914.
47. ↑ Barnabe, Renato Campanarut; Guimarães, Marcelo Alcindo de Barros Vaz; Oliveira, Cláudio Alvarenga de; Barnabe, Alexandre Hyppolito (2002). "Analysis of some normal parameters of the spermogram of captive capuchin monkeys (*Cebus apella Linnaeus, 1758*)". *Brazilian Journal of Veterinary Research and Animal Science* **39**. doi:10.1590/S1413-95962002000600010.
48. ↑ Young WJ, Merz T, Ferguson-Smith MA, Johnston AW (June 1960). "Chromosome number of the chimpanzee, *Pan troglodytes*". *Science* **131**: 1672–3. doi:10.1126/science.131.3414.1672. PMID 13846659.
49. ↑ Jump up to: **a b** <http://resources.metapress.com/pdf-preview.axd?code=3180kk1kk0873012&size=largest>
50. ↑ Lindblad-Toh K, Wade CM, Mikkelsen TS, et al. (December 2005). "Genome sequence, comparative analysis and haplotype structure of the domestic dog". *Nature* **438** (7069): 803–19. doi:10.1038/nature04338. PMID 16341006.
51. ↑ <http://www.ncbi.nlm.nih.gov/genome/guide/dog/>
52. ↑ Guttenbach M, Nanda I, Feichtinger W, Masabanda JS, Griffin DK, Schmid M (2003). "Comparative chromosome painting of chicken autosomal paints 1–9 in nine different bird species". *Cytogenetics and Genome Research* **103** (1–2): 173–84. doi:10.1159/000076309. PMID 15004483.
53. ↑ "Drosophila Genome Project". National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd>ShowDetailView&TermToSearch=9554>. Retrieved 2009-04-14.
54. ↑ T.J. Robinson, F. Yang, W.R. Harrison (2002). "Chromosome

- painting refines the history of genome evolution in hares and rabbits (order Lagomorpha)". *Cytogenetics and Genetic Research* **96**: 223–227. doi:10.1159/000063034. PMID 12438803.
[http://content.karger.com/ProdukteDB/
produkte.asp?Aktion=ShowAbstract&ArtikelNr=63034&Ausgabe
e=228416&ProduktNr=224037](http://content.karger.com/ProdukteDB/produkte.asp?Aktion=ShowAbstract&ArtikelNr=63034&Ausgabe=228416&ProduktNr=224037).
55. ↑ "Rabbits, Hares and Pikas. Status Survey and Conservation Action Plan". pp. 61–94.
[http://wildlife1.wildlifeinformation.org/s/00Ref/
BooksContents/b605.htm](http://wildlife1.wildlifeinformation.org/s/00Ref/BooksContents/b605.htm).
56. ↑ "Human Genome Project". National Center for Biotechnology Information. [http://www.ncbi.nlm.nih.gov/sites/
entrez?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9558](http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=9558). Retrieved 2009-04-29.
57. ↑ Jump up to: ^a ^b Crosland, M.W.J., Crozier, R.H. (1986). "Myrmecia pilosula, an ant with only one pair of chromosomes". *Science* **231** (4743): 1278. doi:10.1126/science.231.4743.1278. PMID 17839565.
58. ↑ Jump up to: ^a ^b Francesco Giannelli; Hall, Jeffrey C.; Dunlap, Jay C.; Friedmann, Theodore (1999). *Advances in Genetics, Volume 41 (Advances in Genetics)*. Boston: Academic Press. p. 2. ISBN 0-12-017641-6.
59. ↑ Perelman PL, Graphodatsky AS, Dragoo JW, Serdyukova NA, Stone G, Cavagna P, Menotti A, Nie W, O'Brien PC, Wang J, Burkett S, Yuki K, Roelke ME, O'Brien SJ, Yang F, Stanyon R (2008). "Chromosome painting shows that skunks (Mephitidae, Carnivora) have highly rearranged karyotypes". *Chromosome Res.* **16** (8): 1215–31. doi:10.1007/s10577-008-1270-2. PMID 19051045.
60. ↑ Hosseini S-J, Elahi E, Raie RM (2004). "The Chromosome Number of the Persian Gulf Shrimp *Penaeus semisulcatus*". *Iranian Int. J. Sci* **5** (1): 13–23.
61. ↑ "First of six chromosomes sequenced in *Dictyostelium discoideum*". Genome News Network.
http://www.genomenewsnetwork.org/articles/07_02/

- dictyostelium.shtml. Retrieved 2009-04-29.
- 62. ↑ Biggers JD, Fritz HI, Hare WC, McFeely RA (June 1965). “Chromosomes of American Marsupials”. *Science* **148**: 1602–3. doi:10.1126/science.148.3677.1602. PMID 14287602.
 - 63. ↑ Abrams, L. (1951). *Illustrated Flora of the Pacific States*. Volume 3.. Stanford University Press. pp. 866.
 - 64. ↑ Stance, C. (1997). *New Flora of the British Isles*. Second Edition.. Cambridge, UK. pp. 1130.
 - 65. ↑ Liang, X; Bing, W. (April 2004). “[Karyotype analysis of *Physalis pubescens* chromosome] (article in Chinese)”. *Zhong Yao Cai*. **27** (4): 238–239.

II.

The cell cycle, or cell-division cycle (cdc), is the series of events that takes place in a cell leading to its division and duplication. In cells without a nucleus (prokaryotic), the cell cycle occurs via a process termed binary fission. In cells with a nucleus (eukaryotes), the cell cycle can be divided in two brief periods: interphase—during which the cell grows, accumulating nutrients needed for mitosis and duplicating its DNA—and the mitosis (M) phase, during which the cell splits itself into two distinct cells, often called “daughter cells”. The cell-division cycle is a vital process by which a single-celled fertilized egg develops into a mature organism, as well as the process by which hair, skin, blood cells, and some internal organs are renewed.^[1]



Schematic of the cell cycle. outer ring: I=Interphase, M=Mitosis; inner ring: M=Mitosis; G₁=Gap phase 1; S=Synthesis; G₂=Gap phase 2. The duration of mitosis in relation to the other phases has been exaggerated in this diagram

Contents

- 1 Phases of cell division
- 2 G₀ phase
- 3 G₁ phase
- 4 S phase
- 5 G₂ phase
- 6 Mitosis

- 7 Cyclins
 - 7.1 Types of Cyclins
- 8 Cyclin dependent kinases (CDKs)
 - 8.1 Functions of cyclin and CDKs
- 9 Disregulation of cell cycle
- 10 Cell cycle checkpoints
 - 10.1 The G1/S checkpoint
 - 10.2 The G2/M checkpoint
- 11 Metaphase-anaphase checkpoint
- 12 Cell division in fission yeast
- 13 Facts to be remembered
- 14 References

Phases of cell division

The cell cycle consists of four distinct phases: G1 (Gap1) phase, S phase (synthesis), G2 (Gap2) phase (collectively known as interphase) and M phase (mitosis). M (mitosis) phase is itself composed of two tightly coupled processes: mitosis, in which the cell's chromosomes are divided between the two daughter cells, and cytokinesis, in which the cell's cytoplasm divides in half forming distinct cells. Activation of each phase is dependent on the proper progression and completion of the previous one. Cells that have temporarily or reversibly stopped dividing are said to have entered a state of quiescence called G0 phase.^[1]

Go phase

The G0 phase is a period in the cell cycle in which cells exist in a quiescent state. G0 phase is viewed as either an extended G1

phase, where the cell is neither dividing nor preparing to divide, or a distinct quiescent stage that occurs outside of the cell cycle. G₀ is sometimes referred to as a “post-mitotic” state, since cells in G₀ are in a non-dividing phase outside of the cell cycle. Some types of cells, such as nerve and heart muscle cells, become post-mitotic when they reach maturity (i.e., when they are terminally differentiated) but continue to perform their main functions for the rest of the organism’s life. Multinucleated muscle cells that do not undergo cytokinesis are also often considered to be in the G₀ stage. On occasion, a distinction in terms is made between a G₀ cell and a ‘post-mitotic’ cell (e.g., heart muscle cells and neurons), which will never enter the G₁ phase, whereas other G₀ cells may. ^[2] G₀ cells are normally referred to by scientists as “P53-Cells”.

G₁ phase

The first phase of interphase is G₁ phase, from the end of the previous Mitosis phase until the beginning of DNA replication is called G₁ (G indicating gap). It is also called the growth phase. During this phase the biosynthetic activities of the cell, which had been considerably slowed down during M phase, resume at a high rate. This phase is marked by synthesis of various enzymes that are required in S phase, mainly those needed for DNA replication. Duration of G₁ is highly variable, even among different cells of the same species.^[1]

S phase

Initiation of DNA replication is indication of S phase; when it is complete, all of the chromosomes have been replicated, at this time each chromosome has two (sister) chromatids. Thus, during this

phase, the amount of DNA in the cell has effectively doubled, though the ploidy of the cell remains the same. Rates of RNA transcription and protein synthesis are very low during this phase. **An exception to this is production of histone protein, which mostly occurs during the S phase.**^[1]

G₂ phase

After S phase or replication cell then enters the G₂ phase, which lasts until the cell enters mitosis. Again, significant biosynthesis occurs during this phase, mainly involving the production of microtubules, which are required during the process of mitosis. Inhibition of protein synthesis during G₂ phase prevents the cell from undergoing mitosis.

Analysis of Cell cycle

Cell cycle analysis is a method in cell biology that employs flow cytometry to distinguish cells in different phases of the cell cycle. Before analysis, the cells are permeabilised and treated with a fluorescent dye that stains DNA quantitatively, usually propidium iodide (PI). The fluorescence intensity of the stained cells at certain wavelengths will therefore correlate with the amount of DNA they contain. As the DNA content of cells duplicates during the S phase of the cell cycle, the relative amount of cells in the G₀ phase and G₁ phase (before S phase), in the S phase, and in the G₂ phase and M phase (after S phase) can be determined, as the fluorescence of cells in the G₂/M phase will be twice as high as that of cells in the G₀/G₁ phase. Cell cycle anomalies can be symptoms for various kinds of cell damage, for example DNA damage, which cause the cell to interrupt the cell cycle at certain checkpoints to prevent transformation into a cancer cell (carcinogenesis). Other possible reasons for anomalies include lack of nutrients, for example after serum deprivation. Cell cycle analysis was first described in 1969 at Los Alamos Scientific Laboratory by a group from the University

of California[1], using the Feulgen staining technique. The first protocol for cell cycle analysis using propidium iodide staining was presented in 1975 by Awtar Krishan from Harvard Medical School and is still widely cited today.^[3]

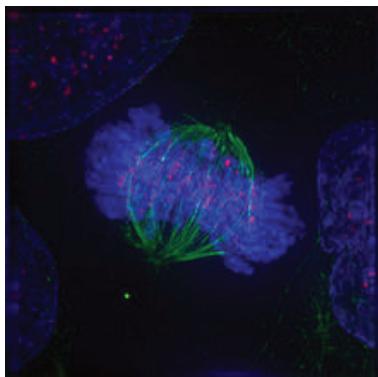
Mitosis

Mitosis is the process by which a eukaryotic cell separates the chromosomes in its nucleus into two identical sets in two nuclei. It is generally followed immediately by cytokinesis, which divides the nuclei, cytoplasm, organelles and cell membrane into two cells containing roughly equal shares of these cellular components. Mitosis and cytokinesis together define the mitotic (M) phase of the cell cycle – the division of the mother cell into two daughter cells, genetically identical to each other and to their parent cell. This accounts for approximately 10% of the cell cycle. Mitosis occurs exclusively in eukaryotic cells, but occurs in different ways in different species. For example, animals undergo an “open” mitosis, where the nuclear envelope breaks down before the chromosomes separate, while fungi such as *Aspergillus nidulans* and *Saccharomyces cerevisiae* (yeast) undergo a “closed” mitosis, where chromosomes divide within an intact cell nucleus. Prokaryotic cells, which lack a nucleus, divide by a process called binary fission. The process of mitosis is complex and highly regulated. The sequence of events is divided into phases, corresponding to the completion of one set of activities and the start of the next. These stages are prophase, prometaphase, metaphase, anaphase and telophase. During the process of mitosis the pairs of chromosomes condense and attach to fibers that pull the sister chromatids to opposite sides of the cell. The cell then divides in cytokinesis, to produce two identical daughter cells. Because cytokinesis usually occurs in conjunction with mitosis, “mitosis” is often used interchangeably with “M phase”. However, there are many cells where mitosis and

cytokinesis occur separately, forming single cells with multiple nuclei. This occurs most notably among the fungi and slime moulds, but is found in various different groups. Even in animals, cytokinesis and mitosis may occur independently, for instance during certain stages of fruit fly embryonic development. Errors in mitosis can either kill a cell through apoptosis or cause mutations that may lead to cancer.^[4]

Prophase

Prophase, from the ancient Greek pro (before) and phase (stage), is a stage of mitosis in which the chromatin condenses (it becomes shorter and fatter) into a highly ordered structure called a chromosome in which the chromatin becomes visible. This process, called chromatin condensation, is mediated by the condensin complex. Since the genetic material has been duplicated in an earlier phase of the cell cycle, there are two identical copies of each chromosome in the cell. Identical chromosomes, called sister chromatids, are attached to each other at a DNA element present on every chromosome called the centromere. During prophase, giemsa staining can be applied to elicit G-banding in chromosomes. Prophase accounts for approximately 3% of the cell cycle's duration. An important organelle in mitosis is the centrosome, the microtubule organizing center in metazoans. During prophase, the two centrosomes, which replicate independently of mitosis, have their microtubule-activity increased due to the recruitment of γ -tubulin. The centrosomes will be pushed apart to opposite ends of the cell nucleus by the action of molecular motors acting on the microtubules. The nuclear envelope breaks

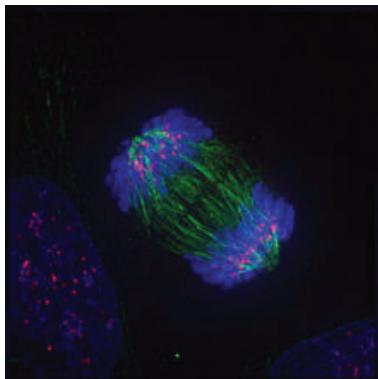


Immunofluorescent image of a cell in metaphase showing microtubules in green, chromosomes (DNA) in blue, and kinetochores in pink

down to allow the microtubules to reach the kinetochores on the chromosomes, marking the end of prophase. Prometaphase, the next step of mitosis, will see the chromosome being captured by the microtubules.^[5]

Prophase in plant cells

In this first phase of mitosis, plant cells undergo a series of changes that is called puberty. In highly vacuolated plant cells, the contractile vacuole has to migrate into the center of the cell before mitosis can begin. This is achieved during the G2 phase of the cell cycle. A transverse sheet of cytoplasm bisects the cell along the future plane of cell division. Prophase in plant cells is preceded by a stage only found in plants, the formation of a ring of microtubules and actin filaments underneath the plasma membrane around the equatorial plane of the future mitotic spindle and predicting the position of cell plate fusion during telophase. During telophase in animal cells, a cleavage furrow forms. The preprophase band disappears during nuclear envelope disassembly and spindle formation in prometaphase despite contrary belief. The cells of higher plants lack centrioles. Instead, the nuclear envelope serves as a microtubule organising center. Spindle microtubules aggregate on the surface of the nuclear envelope during preprophase and prophase, forming the prophase spindle.^[5]



A cell during anaphase.

Metaphase

Metaphase, from the ancient Greek meta (between) and phase (stage), is a stage of mitosis in the eukaryotic cell cycle in which condensed & highly coiled chromosomes, carrying genetic information, align in the middle of the cell before being separated into each of the two daughter cells. Metaphase accounts for

approximately 4% of the cell cycle's duration. Preceded by events in prometaphase and followed by anaphase, microtubules formed in prophase have already found and attached themselves to kinetochores in metaphase. The centromeres of the chromosomes convene themselves on the metaphase plate (or equatorial plate), an imaginary line that is equidistant from the two centrosome poles. This even alignment is due to the counterbalance of the pulling powers generated by the opposing kinetochores, analogous to a tug of war between equally strong people. In certain types of cells, chromosomes do not line up at the metaphase plate and instead move back and forth between the poles randomly, only roughly lining up along the middleline. Early events of metaphase can coincide with the later events of prometaphase, as chromosomes with connected kinetochores will start the events of metaphase individually before other chromosomes with unconnected kinetochores that are still lingering in the events of prometaphase. One of the cell cycle checkpoints occurs during prometaphase and metaphase. Only after all chromosomes have become aligned at the metaphase plate, when every kinetochore is properly attached to a bundle of microtubules, does the cell enter anaphase. It is thought that unattached or improperly attached kinetochores generate a signal to prevent premature progression to anaphase, even if most of the kinetochores have been attached and most of the chromosomes have been aligned. Such a signal creates the mitotic spindle checkpoint. This would be accomplished by regulation of the anaphase-promoting complex, securin, and separase.^[6]

Anaphase (ana (up) and phase (stage))

Anaphase begins abruptly with the regulated triggering of the metaphase-to-anaphase transition and accounts for approximately 1% of the cell cycle's duration. At this point, anaphase begins. This terminate activity by cleaving and inactivating the M-phase cyclin required for the function of M-phase cyclin dependent kinases (M-Cdks). It also cleaves securin, a protein that inhibits the protease known as separase. Separase then cleaves cohesin, a protein responsible for holding sister chromatids together. During early

anaphase (or Anaphase A), the chromatids abruptly separate and move toward the spindle poles. This is achieved by the shortening of spindle microtubules, with forces mainly being exerted at the kinetochores. anaphase is when the chromatids separate from each other and move to opposite ends of the cell When the chromatids are fully separated, late anaphase (or Anaphase B) begins. This involves the polar microtubules elongating and sliding relative to each other to drive the spindle poles to opposite ends of the cell. Anaphase B drives the separation of sister centrosomes to opposite poles through three forces. Kinesin proteins that are attached to polar microtubules push the microtubules past one another. A second force involves the pulling of the microtubules by cortex-associated cytosolic dynein. The third force for chromosome separation involves the lengthening of the polar microtubules at their plus ends. These two processes were originally distinguished by their different sensitivities to drugs, and they are mechanically distinct. Early anaphase (Anaphase A) involves the shortening of kinetochore microtubules by depolymerization at their plus ends. During this process, a sliding collar allows chromatid movement. No motor protein is involved, as ATP depletion does not inhibit early anaphase. Late anaphase (Anaphase B) involves both the elongation of overlapping microtubules and the use of two distinct sets of motor proteins: one pulls overlapping microtubules past each other, and the other pulls astral microtubules that have attached to the cell cortex. The contributions of early anaphase and late anaphase to anaphase as a whole vary by cell type. In mammalian cells, late anaphase follows shortly after early anaphase and extends the spindle to approximately twice its metaphase length; by contrast, yeast and certain protozoans use late metaphase as the main means of chromosome separation and, in the process, can extend their spindles to up to 15 times the metaphase length.^[7]

Cyclins

Cyclins are a Group of proteins that control the progression of cells through the cell cycle by activating Cyclin-dependent kinase (Cdk) enzymes. Cyclins were discovered by **R. Timothy Hunt** in 1982 while studying the cell cycle of sea urchins.

Types of Cyclins

There are several different cyclins that are active in different parts of the cell cycle and that cause the Cdk to phosphorylate different substrates.

There are two groups of cyclins:

G1/S cyclins – These cyclins are essential for the control of the cell cycle at the G1/S transition, Cyclin A / CDK2 – active in S phase. Cyclin D / CDK4, Cyclin D / CDK6, and Cyclin E / CDK2 – regulates transition from G1 to S phase.

G2/M cyclins – essential for the control of the cell cycle at the G2/M transition (mitosis). G2/M cyclins accumulate steadily during G2 and are abruptly destroyed as cells exit from mitosis (at the end of the M-phase). Cyclin B / CDK1 – regulates progression from G2 to M phase.

There are also several “orphan” cyclins for which no Cdk partner has been identified. For example, cyclin F is an orphan cyclin that is essential for G2/M transition.^[8]

Cyclin dependent kinases (CDKs)

CDKs are a family of protein kinases. CDKs are present in all known eukaryotes, and their regulatory function in the cell cycle has been

evolutionarily conserved. CDKs are also involved in regulation of transcription, mRNA processing, and the differentiation of nerve cells. One interesting fact is that, yeast cells can proliferate normally when their CDK gene has been replaced with the homologous human gene. CDKs are relatively small proteins, with molecular weights ranging from 34 to 40 kDa, and contain little more than the kinase domain. CDK binds to a regulatory protein called a cyclin. Without cyclin, CDK has little kinase activity, only the cyclin-CDK complex is an active kinase. CDKs phosphorylate their substrates on serines and threonines, so they are serine-threonine kinases. The consensus sequence for the phosphorylation site in the amino acid sequence of a CDK substrate is **[S/T*]PX[K/R]**, where S/T* is the phosphorylated serine or threonine, P is proline, X is any amino acid, K is lysine, and R is arginine.^[9]

Table : Cyclin-dependent kinases that control the cell cycle in model organisms.^[10]

Species	Name	Original name	Size (amino acids)	Function
Saccharomyces cerevisiae	Cdk1	Cdc28	298	All cell-cycle stages
Schizosaccharomyces pombe	Cdk1	Cdc2	297	All cell-cycle stages
Drosophila melanogaster	Cdk1	Cdc2	297	M
	Cdk2	Cdc2c	314	G1/S, S, possibly M
	Cdk4	Cdk4/6	317	G1, promotes growth
Xenopus laevis	Cdk1	Cdc2	301	M
	Cdk2		297	S, possibly M
Homo sapiens	Cdk1	Cdc2	297	M
	Cdk2		298	G1, S, possibly M
	Cdk4		301	G1
	Cdk6		326	G1

Functions of cyclin and CDKs

Two key classes of regulatory molecules, cyclins and cyclin-dependent kinases (CDKs), determine a cell's progress through the cell cycle. Leland H. Hartwell, R. Timothy Hunt, and Paul M. Nurse won the 2001 Nobel Prize in Physiology or Medicine for their discovery of these central molecules. Many of the genes encoding cyclins and CDKs are conserved among all eukaryotes, but in general more complex organisms have more elaborate cell cycle control systems that incorporate more individual components. Many of the relevant genes were first identified by studying yeast, especially *Saccharomyces cerevisiae*; genetic nomenclature in yeast dubs many these genes cdc (for "cell division cycle") followed by an identifying number, e.g., cdc25 or cdc20.

Cyclins form the regulatory subunits and CDKs the catalytic subunits of an activated heterodimer; cyclins have no catalytic activity and CDKs are inactive in the absence of a partner cyclin. When activated by a bound cyclin, CDKs perform a common biochemical reaction called phosphorylation that activates or inactivates target proteins to orchestrate coordinated entry into the next phase of the cell cycle. Different cyclin-CDK combinations determine the downstream proteins targeted. CDKs are constitutively expressed in cells whereas cyclins are synthesised at specific stages of the cell cycle, in response to various molecular signals.

Upon receiving a pro-mitotic extracellular signal, G1 cyclin-CDK complexes become active to prepare the cell for S phase, promoting the expression of transcription factors that in turn promote the expression of S cyclins and of enzymes required for DNA replication. The G1 cyclin-CDK complexes also promote the degradation of molecules that function as S phase inhibitors by targeting them for ubiquitination. Once a protein has been ubiquitinated, it is targeted for proteolytic degradation by the proteasome. Active S cyclin-CDK complexes phosphorylate proteins that make up the pre-replication

complexes assembled during G1 phase on DNA replication origins. The phosphorylation serves two purposes: to activate each already-assembled pre-replication complex, and to prevent new complexes from forming. This ensures that every portion of the cell's genome will be replicated once and only once. The reason for prevention of gaps in replication is fairly clear, because daughter cells that are missing all or part of crucial genes will die. However, for reasons related to gene copy number effects, possession of extra copies of certain genes is also deleterious to the daughter cells. Mitotic cyclin-CDK complexes, which are synthesized but inactivated during S and G2 phases, promote the initiation of mitosis by stimulating downstream proteins involved in chromosome condensation and mitotic spindle assembly. A critical complex activated during this process is a ubiquitin ligase known as the anaphase-promoting complex (APC), which promotes degradation of structural proteins associated with the chromosomal kinetochore. APC also targets the mitotic cyclins for degradation, ensuring that telophase and cytokinesis can proceed. Interphase: Interphase generally lasts at least 12 to 24 hours in mammalian tissue. During this period, the cell is constantly synthesizing RNA, producing protein and growing in size. By studying molecular events in cells, scientists have determined that interphase can be divided into 4 steps: Gap 0 (G0), Gap 1 (G1), S (synthesis) phase, Gap 2 (G2).

Cyclin D is the first cyclin produced in the cell cycle, in response to extracellular signals (e.g. growth factors). Cyclin D binds to existing CDK4, forming the active cyclin D-CDK4 complex. Cyclin D-CDK4 complex in turn phosphorylates the retinoblastoma susceptibility protein (Rb). The hyperphosphorylated Rb dissociates from the E2F/DP1/Rb complex (which was bound to the E2F responsive genes, effectively "blocking" them from transcription), activating E2F. Activation of E2F results in transcription of various genes like cyclin E, cyclin A, DNA polymerase, thymidine kinase, etc. Cyclin E thus produced binds to CDK2, forming the cyclin E-CDK2 complex, which pushes the cell from G1 to S phase (G1/S

transition). Cyclin B along with cdc2 (cdc2 – fission yeasts (CDK1 – mammalia)) forms the cyclin B-cdc2 complex, which initiates the G2/M transition. Cyclin B-cdc2 complex activation causes breakdown of nuclear envelope and initiation of prophase, and subsequently, its deactivation causes the cell to exit mitosis.^[1]

Disregulation of cell cycle

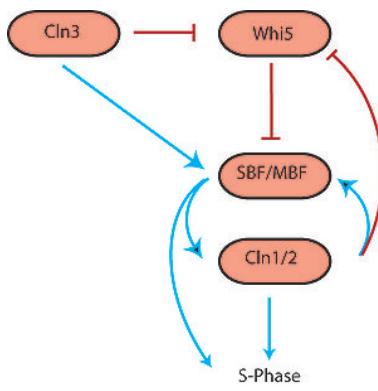
A disregulation of the cell cycle components may lead to tumor formation. As mentioned above, some genes like the cell cycle inhibitors, RB, p53 etc., when they mutate, may cause the cell to multiply uncontrollably, forming a tumor. Although the duration of cell cycle in tumor cells is equal to or longer than that of normal cell cycle, the proportion of cells that are in active cell division (versus quiescent cells in G0 phase) in tumors is much higher than that in normal tissue. Thus there is a net increase in cell number as the number of cells that die by apoptosis or senescence remains the same. The cells which are actively undergoing cell cycle are targeted in cancer therapy as the DNA is relatively exposed during cell division and hence susceptible to damage by drugs or radiation. This fact is made use of in cancer treatment; by a process known as debulking, a significant mass of the tumor is removed which pushes a significant number of the remaining tumor cells from G0 to G1 phase (due to increased availability of nutrients, oxygen, growth factors etc.). Radiation or chemotherapy following the debulking procedure kills these cells which have newly entered the cell cycle. The fastest cycling mammalian cells in culture, crypt cells in the intestinal epithelium, have a cycle time as short as 9 to 10 hours. Stem cells in resting mouse skin may have a cycle time of more than 200 hours. Most of this difference is due to the varying length of G1, the most variable phase of the cycle. M and S do not vary much. In general, cells are most radiosensitive in late M and G2 phases and most resistant in late S. For cells with a longer cell cycle time and

a significantly long G1 phase, there is a second peak of resistance late in G1. The pattern of resistance and sensitivity correlates with the level of sulphydryl compounds in the cell. Sulphydryls are natural radioprotectors and tend to be at their highest levels in S and at their lowest near mitosis.^[1]

Cell cycle checkpoints

The G₁/S checkpoint

The G₁/S transition, more commonly known as the Start checkpoint in budding yeast (the restriction point in other organisms) regulates cell cycle commitment. At this checkpoint, cells either arrest before DNA replication (due to limiting nutrients or a pheromone signal), prolong G₁ (size control), or begin replication and progress through the rest of the cell cycle. The G₁/S regulatory network or regulon in budding yeast includes the G₁ cyclins Cln1, Cln2 and Cln3, Cdc28 (Cdk1), the transcription factors SBF and MBF, and the transcriptional inhibitor Whi5.^[11] Cln3 interacts with Cdk1 to initiate the sequence of events by phosphorylating a large number of targets, including SBF, MBF and Whi5. Phosphorylation of Whi5 causes it to translocate out of the nucleus, preventing it from inhibiting SBF and MBF. Active SBF/MBF drive the G₁/S transition by turning on the B-type cyclins and initiating DNA replication, bud formation and spindle body duplication. Moreover, SBF/MBF drives



expression of Cln1 and Cln2, which can also interact with Cdk1 to promote phosphorylation of its targets.

This G1/S switch was initially thought to function as a linear sequence of events starting with Cln3 and ending in S phase.^[12] However, the observation that any one of the Clns was sufficient to activate the regulon indicated that Cln1 and Cln2 might be able to engage positive feedback to activate their own transcription. This would result in a continuously accelerating cycle that could act as an irreversible bistable trigger. Skotheim et al. used single-cell measurements in budding yeast to show that this positive feedback does indeed occur. A small amount of Cln3 induces Cln1/2 expression and then the feedback loop takes over, leading to rapid and abrupt exit of Whi5 from the nucleus and consequently coherent expression of G1/S regulon genes. In the absence of coherent gene expression, cells take longer to exit G1 and a significant fraction even arrest before S phase, highlighting the importance of positive feedback in sharpening the G1/S switch.

The G1/S cell cycle checkpoint controls the passage of eukaryotic cells from the first gap phase, G1, into the DNA synthesis phase, S. In this switch in mammalian cells, there are two cell cycle kinases that help to control the checkpoint: cell cycle kinases CDK4/6-cyclin D and CDK2-cyclin E. The transcription complex that includes Rb and E2F is important in controlling this checkpoint. In the first gap phase, the Rb-HDAC repressor complex binds to the E2F-DP1 transcription factors, therefore inhibiting the downstream transcription. The phosphorylation of Rb by CDK4/6 and CDK2 dissociates the Rb-repressor complex and serves as an on/off switch for the cell cycle. Once Rb is phosphorylated, the inhibition is released on the E2F transcriptional activity. This allows for the transcription of S phase genes encoding for proteins that amplify the G1 to S phase switch.^[13]

Many different stimuli apply checkpoint controls including TGF β , DNA damage, contact inhibition, replicative senescence, and growth factor withdrawal. The first four act by inducing members of the INK4 or Kip/Cip families of cell cycle kinase inhibitors. TGF β

inhibits the transcription of Cdc25A, a phosphatase that activates the cell cycle kinases, and growth factor withdrawal activates GSK3b, which phosphorylates cyclin D. This leads to its rapid ubiquitination.^[14]

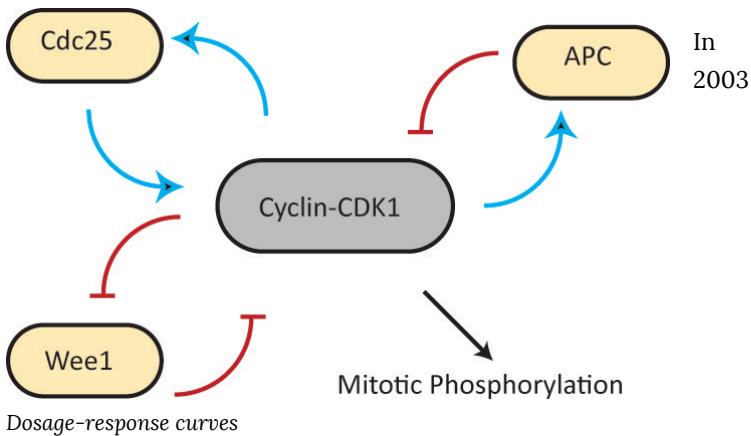
The G₂/M checkpoint

This transition is commenced by E2F-mediated transcription of cyclin A, forming the cyclin A-Cdk2 complex. This is useful in regulating events in prophase. In order to proceed past prophase, the cyclin B-Cdk1 complex (first discovered as MPF or M-phase promoting factor) is activated by Cdc 25, a protein phosphatase¹. As mitosis starts, the nuclear envelope disintegrates, chromosomes condense and become visible, and the cells prepares for division. The Cyclin B-Cdk1 activation results in nuclear envelope breakdown, which is a characteristic of the initiation of mitosis. It is evident that the cyclin A and B complexes with Cdks help regulate mitotic events at the G₂/M transition.^[15]

As mentioned above, entry into mitosis is controlled by the Cyclin B-Cdk1 complex (first discovered as MPF or M-phase promoting factor; Cdk1 is also known as Cdc2 in fission yeast and Cdc28 in budding yeast). This complex forms an element of an interesting regulatory circuit in which Cdk1 can phosphorylate and activate its activator, the phosphatase Cdc25 (positive feedback), and phosphorylate and inactivate its inactivator, the kinase Wee1 (double-negative feedback). It was suggested that this circuit could act as a bistable trigger^[16] with one stable steady state in G₂ (Cdk and Cdc25 off, Wee1 on) and a second stable steady state in M phase (Cdk and Cdc25 active, Wee1 off). Once cells are in mitosis, Cyclin B-Cdk1 activates the Anaphase-promoting complex (APC), which in turn inactivates Cyclin B-Cdk1 by degrading Cyclin B, eventually leading to exit from mitosis. Coupling the bistable Cdk1 response function to the negative feedback from the APC could generate

what is known as a relaxation oscillator,^[17] with sharp spikes of Cdk1 activity triggering robust mitotic cycles. However, in a relaxation oscillator, the control parameter moves slowly relative to the system's response dynamics which may be an accurate representation of mitotic entry, but not necessarily mitotic exit.

It is necessary to inactivate the cyclin B-Cdk1 complex in order to exit the mitotic stage of the cell cycle. The cells can then return to the first gap phase G1 and wait until the cycle proceeds yet again.



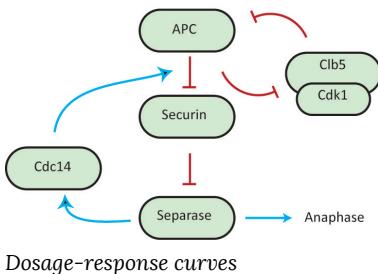
Pomerening et al. provided strong evidence for this hypothesis by demonstrating hysteresis and bistability in the activation of Cdk1 in the cytoplasmic extracts of Xenopus oocytes.^[17] They first demonstrated a discontinuous sharp response of Cdk1 to changing concentrations of non-destructible Cyclin B (to decouple the Cdk1 response network from APC-mediated negative feedback). However, such a response would be consistent with both a monostable, ultrasensitive transition and a bistable transition. To distinguish between these two possibilities, they measured the steady-state levels of active Cdk1 in response to changing cyclin levels, but in two separate experiments, one starting with an interphase extract and one starting with an extract already in mitosis. At intermediate concentrations of cyclin they found two steady-state concentrations of active Cdk1. Which of the two steady states was occupied depended on the history of the system, i.e. whether they

started with interphase or mitotic extract, effectively demonstrating hysteresis and bistability.

In the same year, Sha et al.^[18] independently reached the same conclusion revealing the hysteretic loop also using *Xenopus laevis* egg extracts. In this article, three predictions of the Novak-Tyson model were tested in an effort to conclude that hysteresis is the driving force for “cell-cycle transitions into and out of mitosis”. The predictions of the Novak-Tyson model are generic to all saddle-node bifurcations. Saddle-node bifurcations are extremely useful bifurcations in an imperfect world because they help describe biological systems which are not perfect. The first prediction was that the threshold concentration of cyclin to enter mitosis is higher than the threshold concentration of cyclin to exit mitosis, and this was confirmed by supplementing cycling egg extracts with non-degradable cyclin B and measuring the activation and inactivation threshold after the addition of cycloheximide (CHX), which is a protein synthesis inhibitor. Furthermore, the second prediction of the Novak-Tyson model was also validated: unreplicated deoxyribonucleic acid, or DNA, increases the threshold concentration of cyclin that is required to enter mitosis. In order to arrive at this conclusion, cytostatic factor released extracts were supplemented with CHX, APH (a DNA polymerase inhibitor), or both, and non-degradable cyclin B was added. The third and last prediction that was tested and proven true in this article was that the rate of Cdc2 activation slows down near the activation threshold concentration of cyclin. These predictions and experiments demonstrate the toggle-like switching behavior that can be described by hysteresis in a dynamical system.^[19]

Metaphase-anaphase checkpoint

In the transition from Spindle checkpoint/metaphase to anaphase, it is crucial that sister chromatids are properly and simultaneously separated to opposite ends of the cell. Separation of sister-chromatids is initially strongly inhibited to prevent premature separation



in late mitosis, but this inhibition is relieved through destruction of the inhibitory elements by the anaphase-promoting complex (APC) once sister-chromatid bi-orientation is achieved. One of these inhibitory elements is securin, which prevents the destruction of cohesin, the complex that holds the sister-chromatids together, by binding the protease separase which targets Scc1, a subunit of the cohesin complex, for destruction. In this system, the phosphatase Cdc14 can remove an inhibitory phosphate from securin, thereby facilitating the destruction of securin by the APC, releasing separase. As shown by Uhlmann et al., during the attachment of chromosomes to the mitotic spindle the chromatids remain paired because cohesion between the sisters prevents separation.^[20] Cohesion is established during DNA replication and depends on cohesin, which is a multisubunit complex composed of Scc1, Scc3, Smc2, and Smc3. In yeast at the metaphase-to-anaphase transition, Scc1 dissociates from the chromosomes and the sister chromatids separate. This action is controlled by the Esp1 protein, which is tightly bound by the anaphase inhibitor Pds1 that is destroyed by the anaphase-promoting complex. In order to verify that Esp1 does play a role in regulating Scc1 chromosome association, cell strains were arrested in G1 with an alpha factor. These cells stayed in arrest during the development. Esp1-1 mutant cells were used and the experiment was repeated, and Scc1 successfully bound to the

chromosomes and remained associated even after the synthesis was terminated. This was crucial in showing that with Esp1, Scc1 is hindered in its ability to become stably associated with chromosomes during G1, and Esp1 can in fact directly remove Scc1 from chromosomes.^[13] It has been shown by Holt et al.^[21] that separase activates Cdc14, which in turn acts on securin, thus creating a positive feedback loop that increases the sharpness of the metaphase to anaphase transition and coordination of sister-chromatid separation.^[21] Holt et al. probed the basis for the effect of positive feedback in securin phosphophorlyation by using mutant ‘securin’ strains of yeast, and testing how changes in the phosphoregulation of securin affects the synchrony of sister chromatid separation. Their results indicate that interfering with this positive securin-separase-cdc14 loop decreases sister chromatid separation synchrony. This positive feedback can hypothetically generate bistability in the transition to anaphase, causing the cell to make the irreversible decision to separate sister-chromatids.

Cell division in fission yeast

The fission yeast is a single-celled fungus with simple, fully characterized genome and a rapid growth rate. It has long since been used in brewing, baking and molecular genetics. *S. pombe* is a rod-shaped cell, approximately 3 μm in diameter, that grows entirely by elongation at the ends. After mitosis, division occurs by the formation of a septum, or cell plate, that cleaves the cell at its midpoint.

The central events of cell reproduction are chromosome duplication, which takes place in S (Synthetic) phase, followed by chromosome segregation and nuclear division (mitosis) and cell division (cytokinesis), which are collectively called M (Mitotic) phase. G1 is the gap between M and S phases, and G2 is the gap

between S and M phases. In the budding yeast, the G₂ phase is particularly extended, and cytokinesis (daughter-cell segregation) does not happen until a new S (Synthetic) phase is launched.

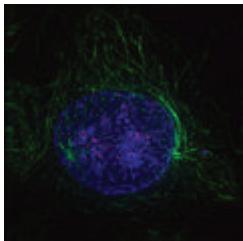
Fission yeast governs mitosis by mechanisms that are similar to those in multicellular animals. It normally proliferates in a haploid state. When starved, cells of opposite mating types (P and M) fuse to form a diploid zygote that immediately enters meiosis to generate four haploid spores. When conditions improve, these spores germinate to produce proliferating haploid cells.^[22]

Facts to be remembered

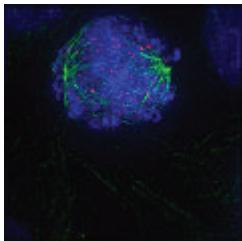
State	Phase	Abbreviation	Description
quiescent/ senescent	Gap 0	G ₀	A resting phase where the cell has left the cycle and has stopped dividing.
	Gap 1	G ₁	Cells increase in size in Gap 1. The G ₁ checkpoint control mechanism ensures that everything is ready for DNA synthesis.
Interphase	Synthesis	S	DNA replication occurs during this phase.
	Gap 2	G ₂	During the gap between DNA synthesis and mitosis, the cell will continue to grow. The G ₂ checkpoint control mechanism ensures that everything is ready to enter the M (mitosis) phase and divide.
Cell division	Mitosis	M	Cell growth stops at this stage and cellular energy is focused on the orderly division into two daughter cells. A checkpoint in the middle of mitosis (Metaphase Checkpoint) ensures that the cell is ready to complete cell division.

Stages of mitosis [4]

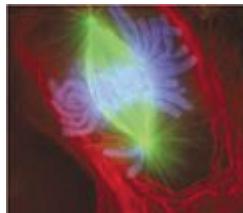
Real mitotic cells can be visualized through the microscope by staining them with fluorescent antibodies and dyes. These light micrographs are included below.



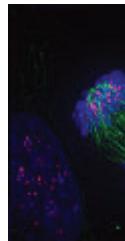
Early prophase:
Nonkinetochore microtubules, shown as green strands, have established a matrix around the degrading nucleus, in blue. The green nodules are the centrosomes.



Early prometaphase:
The nuclear membrane has just degraded, allowing the microtubules to quickly interact with the kinetochores on the chromosomes, which have just condensed.



Late metaphase:
The centrosomes have moved to the poles of the cell and have established the mitotic spindle. The chromosomes, in light blue, have all assembled at the metaphase plate, except for one.



Anaphase
Lengthening nonkinetochore microtubules pull two sets of chromosomes apart.

References

1. ↑ Jump up to: **a b c d e f** Cell cycle
2. ↑ G₀ phase
3. ↑ Cell cycle analysis
4. ↑ Jump up to: **a b** Mitosis
5. ↑ Jump up to: **a b** Prophase
6. ↑ Metaphase
7. ↑ Anaphase
8. ↑ Cyclins
9. ↑ Cyclin-dependent kinase
10. ↑ Morgan, David O. (2007). *The Cell Cycle: Principles of Control*. London: New Science Press, 1st ed.
11. ↑ Skotheim, J.M.; Di Talia, S.; Siggia, E.D.; Cross, F.R. (2008),

- "Positive feedback of G1 cyclins ensures coherent cell cycle entry", *Nature* **454** (7202): 291,
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2606905/>, retrieved 2009-12-11
- 12. ↑ Stuart, D.; Wittenberg, C. (1995), "CLN3, not positive feedback, determines the timing of CLN2 transcription in cycling cells.", *Genes & development* **9** (22): 2780,
<http://genesdev.cshlp.org/cgi/reprint/9/22/2780.pdf>,
retrieved 2009-12-11
 - 13. ↑ Jump up to: **a b** Biochemical switches in the cell cycle
 - 14. ↑ Harper JW. A phosphorylation-driven ubiquitination switch for cell cycle control. *Trends Cell Biol.* 2002 Mar;12(3):104-7.
PMID 11859016
 - 15. ↑ Biochemical switches in the cell cycle
 - 16. ↑ Novak, B.; Tyson, J.J. (1993), "Numerical analysis of a comprehensive model of M-phase control in Xenopus oocyte extracts and intact embryos", *Journal of Cell Science* **106** (4): 1153, <http://jcs.biologists.org/cgi/reprint/106/4/1153.pdf>,
retrieved 2009-12-11
 - 17. ↑ Jump up to: **a b** Pomerening, J. R., E. D. Sontag, et al. (2003). "Building a cell cycle oscillator: hysteresis and bistability in the activation of Cdc2." *Nat Cell Biol* 5(4): 346-351.
 - 18. ↑ Sha, W.; Moore, J.; Chen, K.; Lassaletta, A.D.; Yi, C.S.; Tyson, J.J.; Sible, J.C. (2003), "Hysteresis drives cell-cycle transitions in *Xenopus laevis* egg extracts", *Proceedings of the National Academy of Sciences* **100** (3): 975, <http://www.pnas.org/cgi/content/full/100/3/975>, retrieved 2009-12-11
 - 19. ↑ Cooper, G. (2000), "The Cell: A Molecular Approach.",
retrieved 2010-11-21
 - 20. ↑ Uhlmann F.; Lottspeich F.; Nasmyth K. (1999), "Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesion subunit Scc1," *Nature* **400**: 37-42,
retrieved 2010-9-25
 - 21. ↑ Jump up to: **a b** Holt, L. J., A. N. Krutchinsky, et al. (2008). "Positive feedback sharpens the anaphase switch." *Nature*

454(7202): 353-357.

22. ↑ *Schizosaccharomyces pombe*

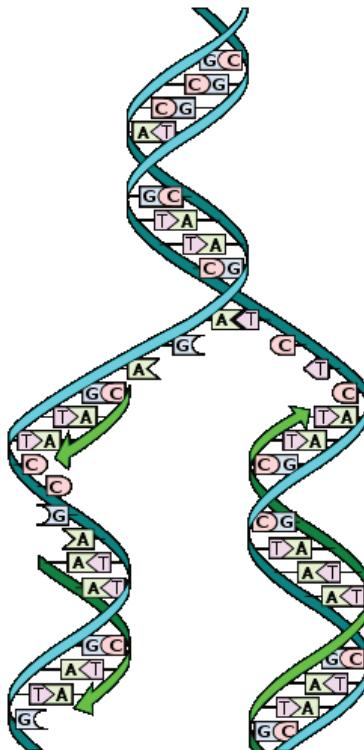
I2.

Genes are made from a long molecule called DNA, which is copied and inherited across generations. DNA is made of simple units that line up in a particular order within this large molecule. The order of these units carries genetic information, similar to how the order of letters on a page carry information. The language used by DNA is called the genetic code, which lets organisms read the information in the genes. This information is the instructions for constructing and operating a living organism. **Deoxyribonucleic acid(DNA):** Deoxyribonucleic acid (/di'ɒksi,raɪbə'nju:kliɪk 'æsɪd/, 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. The structure of DNA was first discovered by **James D. Watson and Francis Crick.** It is the same for all species, comprising two helical chains each coiled round the same axis, each with a pitch of 34

Ångströms (3.4 nanometres) and a radius of 10 Ångströms (1.0 nanometres). 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. 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.^[1]

Contents

- 1 DNA is a genetic material
- 2 Structure of DNA
 - 2.1 Base pairing Of DNA
 - 2.1.1 Purine base
 - 2.1.2 Adenine
 - 2.1.3 Guanine
 - 2.2 Pyrimidine base
 - 2.2.1 Cytosine
 - 2.2.2 Thymine
 - 2.2.3 Uracil
 - 2.3 Nucleosides
 - 2.4 Nucleotide
- 3 Forms of DNA
- 4 Noncoding genomic DNA
- 5 Coiling of DNA
- 6 Histones: The DNA binding protein
 - 6.1 Histone DNA interaction



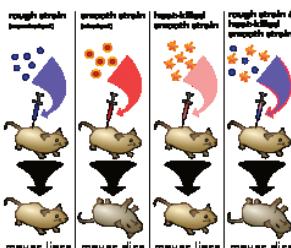
DNA replication. DNA is unwound and nucleotides are matched to make two new strands.

- 7 DNA-binding domains
- 8 DNA sequencing
- 9 Maxam and Gilbert method
- 10 Dideoxynucleotide Chain-termination methods
 - 10.1 Dye-terminator sequencing
 - 10.2 Challenges
 - 10.3 Automation and sample preparation
- 11 Polymerase chain reaction
- 12 Facts to be remembered
- 13 References

DNA is a genetic material

Griffith's experiment was conducted in 1928 by Frederick Griffith, one of the first experiments suggesting that bacteria are capable of transferring genetic information through a process known as transformation.

Griffith used two strains of *Streptococcus pneumoniae* bacteria which infect mice – a type III-S (smooth) and type II-R (rough) strain. The III-S strain covers itself with a polysaccharide capsule that protects it from the host's immune system, resulting in the death of the host, while the II-R strain doesn't have that protective capsule and is defeated by the host's immune system. A German bacteriologist, Fred Neufeld, had discovered the three pneumococcal types (Types I, II, and III) and discovered the Quellung reaction to identify them in vitro. Until Griffith's experiment, bacteriologists believed that the types were fixed and unchangeable, from one generation to another. In this experiment, bacteria from the III-S strain were killed by heat, and their remains were added to II-R strain bacteria. While neither alone harmed the mice, the combination was able to kill its host. Griffith was also able to isolate both live II-R and live III-S strains of pneumococcus from the blood of these dead mice. Griffith concluded that the type II-R had been “transformed” into the lethal III-



Griffith's experiment discovering the “transforming principle” in pneumococcus bacteria.

S strain by a “transforming principle” that was somehow part of the dead III-S

strain bacteria. Today, we know that the “transforming principle” Griffith observed was the DNA of the III-S strain bacteria. While the bacteria had been killed, the DNA had survived the heating process and was taken up by the II-R strain bacteria. The III-S strain DNA contains the genes that form the protective polysaccharide capsule. Equipped with this gene, the former II-R strain bacteria were now protected from the host's immune system and could kill the host. The exact nature of the transforming principle (DNA) was verified in the experiments done by Avery, MacLeod and McCarty and by Hershey and Chase.^[2]

First confirmation:

Alfred Hershey and Martha Chase conducted series of experiments in 1952 by , confirming that DNA was the genetic material, which had first been demonstrated in the 1944 Avery–MacLeod–McCarty experiment. These experiments are known as **Hershey Chase experiments**. The existence of DNA was known to biologists since 1869, most of them assumed that proteins carried the information for inheritance that time. Hershey and Chase conducted their experiments on the T2 phage. The phage consists of a protein shell containing its genetic material. The phage infects a bacterium by attaching to its outer membrane and injecting its genetic material and leaving its empty shell attached to the bacterium. In their first set of experiments, Hershey and Chase labeled the DNA of phages with radioactive Phosphorus-32 (p32) (the element phosphorus is present in DNA but not present in any of the 20 amino acids which are component of proteins). They allowed the phages to infect E. coli, and through several elegant experiments were able to observe the transfer of P32 labeled phage DNA into the cytoplasm of the bacterium. In their second set of

experiments, they labeled the phages with radioactive Sulfur-35 (Sulfur is present in the amino acids cysteine and methionine, but not in DNA). Following infection of *E. coli* they then sheared the viral protein shells off of infected cells using a high-speed blender and separated the cells and viral coats by using a centrifuge. After separation, the radioactive S₃₅ tracer was observed in the protein shells, but not in the infected bacteria, supporting the hypothesis that the genetic material which infects the bacteria was DNA and not protein.^{[3][4]}

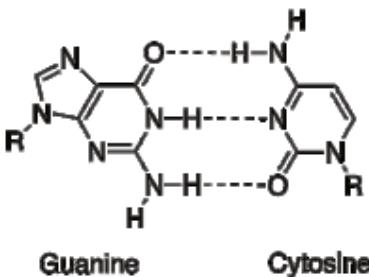
Hershey shared the 1969 Nobel Prize in Physiology or Medicine for his “discoveries concerning the genetic structure of viruses.”



Oswald T. Avery, Colin MacLeod, Maclyn McCarty with Francis Crick and James D Watson ^[5]

Structure of DNA

Two 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 \AA wide and the other, the **minor groove**, is 12 \AA wide. The narrowness of the minor groove means that the edges of the bases are more accessible in the major groove. As a result, proteins like transcription factors that can bind to specific sequences in double-stranded DNA usually make contacts to the sides of the bases exposed in the major groove. This situation varies in unusual conformations of DNA within the cell, but the major and minor grooves are always named to reflect the differences in size that would be seen if the DNA is twisted back into the ordinary B form.

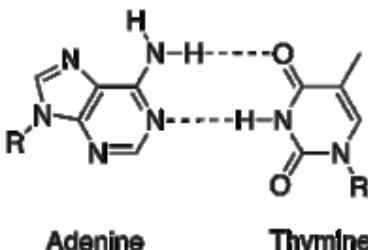


A GC base pair demonstrating three intermolecular hydrogen bonds.

Base pairing Of DNA

Chargaff's rules was given by Erwin Chargaff which state that DNA from any cell of all organisms should have a 1:1 ratio of pyrimidine and purine bases and, more specifically, that the amount of guanine is equal to cytosine and the amount of adenine is equal to thymine. This pattern is found in both strands of the DNA. They were discovered by Austrian chemist Erwin Chargaff.

In molecular biology, two nucleotides on opposite complementary DNA strands that are connected via hydrogen bonds are called a base pair (often abbreviated **bp**). In the canonical Watson-Crick DNA base pairing, Adenine (A) forms a base pair with Thymine (T) and Guanine (G) forms a base pair with Cytosine (C). **In RNA, thymine is replaced by Uracil (U).** Alternate hydrogen bonding patterns, such as the wobble base pair and Hoogsteen base pair, also occur—particularly in RNA—giving rise to complex and functional tertiary structures.^[6]



An AT base pair demonstrating two intermolecular hydrogen bonds.

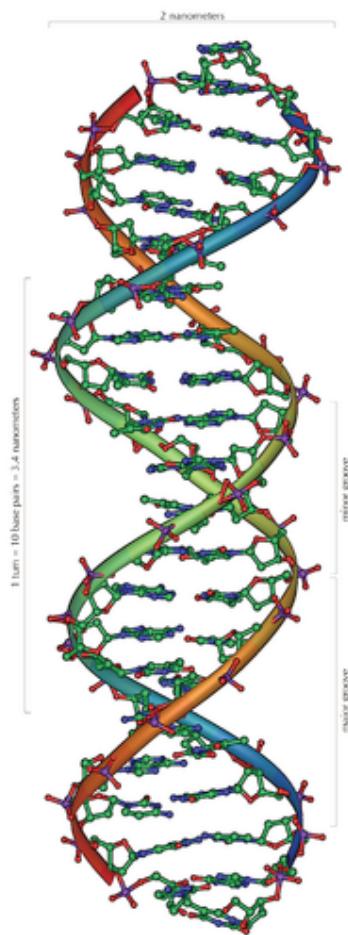
Example

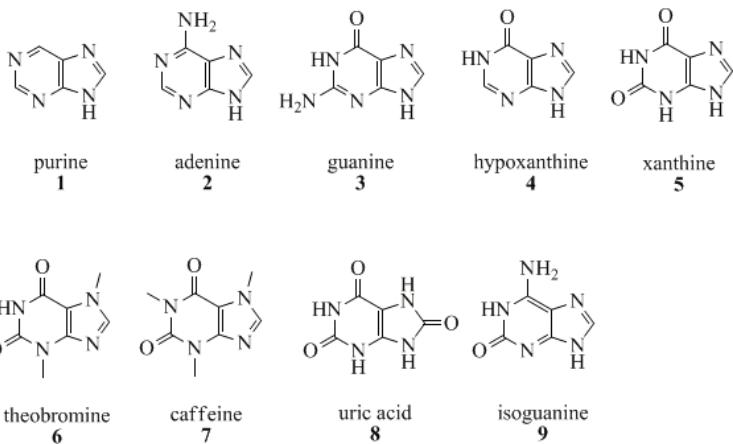
5' CTCGTTGCGCTCTATCG3'
3' GAGCAAACGCGAGATAGC5'

Purine base

The German chemist Emil Fischer in 1884 gave the name 'purine' (purum uricum). He synthesized it for the first time in 1899 by uric acid which had been isolated from kidney stones by Scheele in 1776. Beside from DNA and RNA, purines are also components in a number of other important biomolecules, such as ATP, GTP, cyclic AMP, NADH, and coenzyme A. Purine itself, has not been found in nature, but it can be produced by organic synthesis. A purine is a heterocyclic aromatic organic compound, consisting of a pyrimidine ring fused to an imidazole ring.

Example:





Adenine

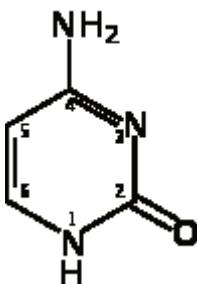
Adenine is one of the two purine nucleobases (the other being guanine) used in forming nucleotides of the nucleic acids (DNA or RNA). In DNA, adenine binds to thymine via two hydrogen bonds to assist in stabilizing the nucleic acid structures. Adenine forms adenosine, a nucleoside, when attached to ribose, and deoxyadenosine when attached to deoxyribose. It forms adenosine triphosphate (ATP), a nucleotide, when three phosphate groups are added to adenosine.

Guanine

Guanine, along with adenine and cytosine, is present in both DNA and RNA, whereas thymine is usually seen only in DNA, and uracil only in RNA. In DNA, guanine is paired with cytosine. With the formula $\text{C}_5\text{H}_5\text{N}_5\text{O}$, guanine is a derivative of purine, consisting of a fused pyrimidine-imidazole ring system with conjugated double bonds.

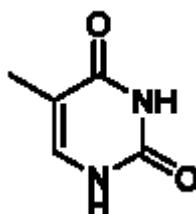
Guanine has two tautomeric forms, the major keto form and rare enol form. It binds to cytosine through three hydrogen bonds. In cytosine, the amino group acts as the hydrogen donor and the C-2 carbonyl and the N-3 amine as the hydrogen-bond acceptors. Guanine has a group at C-6 that acts as the hydrogen acceptor, while the group at N-1 and the amino group at C-2 act as the hydrogen donors.

Pyrimidine base



Cytosine with numbered components.
Methylation occurs on carbon number 5.

Pyrimidine is a heterocyclic aromatic organic compound similar to benzene and pyridine, containing two nitrogen atoms at positions 1 and 3 of the six-member ring. It is isomeric with two other forms of diazine. Three nucleobases found in nucleic acids, **cytosine (C)**, **thymine (T)**, and **uracil (U)**, are pyrimidine derivatives. A pyrimidine has many properties in common with pyridine, as the number of nitrogen atoms in the ring increases the ring pi electrons become less energetic and electrophilic aromatic substitution gets more difficult while nucleophilic aromatic substitution gets easier.



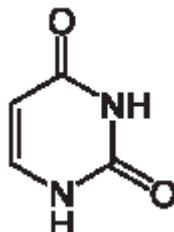
Chemical structure of thymine

An example of the last reaction type is the displacement of the amino group in 2-aminopyrimidine by chlorine and its reverse. Reduction in resonance stabilization of pyrimidines may lead to addition and ring cleavage reactions rather than substitutions. One such manifestation is observed in the Dimroth rearrangement. Compared to pyridine, N-alkylation and N-oxidation is more

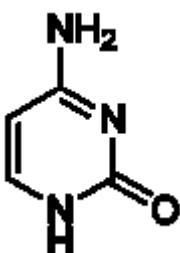
difficult, and pyrimidines are also less basic: The pKa value for protonated pyrimidine is **1.23** compared to **5.30** for pyridine.^[7]

Pyrimidine also is found in meteorites, although scientists still do not know its origin.

Pyrimidine also photolytically decomposes into Uracil under UV light.



Cytosine



Chemical structure of cytosine

Chemical structure of uracil

Cytosine can be found as part of DNA, as part of RNA, or as a part of a nucleotide. As cytidine triphosphate (CTP), it can act as a co-factor to enzymes, and can transfer a phosphate to convert adenosine diphosphate (ADP) to adenosine triphosphate (ATP). The nucleoside of cytosine is cytidine. In DNA and RNA, **cytosine is paired with guanine**. However, it is inherently unstable, and can change into uracil (spontaneous deamination). This can lead to a point mutation if not repaired by the DNA repair enzymes such as uracil glycosylase, which cleaves a uracil in DNA. Cytosine can also be methylated into 5-methylcytosine by an enzyme called DNA methyltransferase or be methylated and hydroxylated to make 5-hydroxymethylcytosine. Active enzymatic deamination of cytosine or 5-methylcytosine by the APOBEC family of cytosine deaminases could have both beneficial and detrimental implications on various cellular processes as well as on organismal evolution. The implications of deamination on 5-hydroxymethylcytosine, on the other hand, remains less understood.^[8]

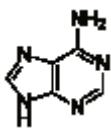
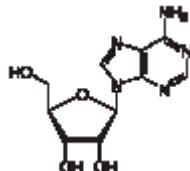
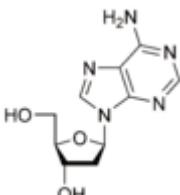
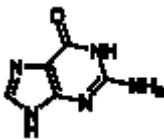
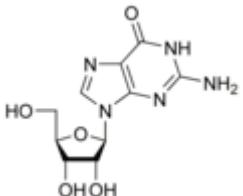
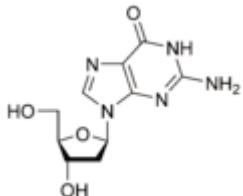
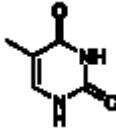
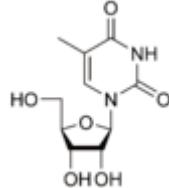
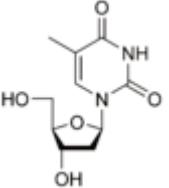
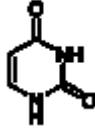
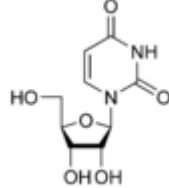
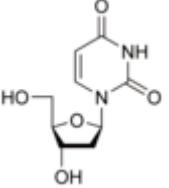
Thymine

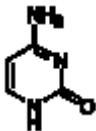
Thymine (T, Thy) is one of the four nucleobases in the nucleic acid of DNA that are represented by the letters G-C-A-T. The others are adenine, guanine, and cytosine. Thymine is also known as 5-methyluracil, a pyrimidine nucleobase. As the name suggests, thymine may be derived by methylation of uracil at the 5th carbon. In RNA, thymine is replaced with uracil in most cases. In DNA, thymine(T) binds to adenine (A) via two hydrogen bonds, thus stabilizing the nucleic acid structures.

Uracil

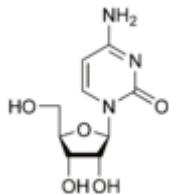
Uracil found in RNA, it base-pairs with adenine and replaces thymine during DNA transcription. Methylation of uracil produces thymine. It turns into thymine to protect the DNA and to improve the efficiency of DNA replication. Uracil can base-pair with any of the bases, depending on how the molecule arranges itself on the helix, but readily pairs with adenine because the methyl group is repelled into a fixed position. Uracil pairs with adenine through hydrogen bonding. Uracil is the hydrogen bond acceptor and can form two hydrogen bonds. Uracil can also bind with a ribose sugar to form the ribonucleoside uridine. When a phosphate attaches to uridine, uridine 5'-monophosphate is produced.

Nucleosides

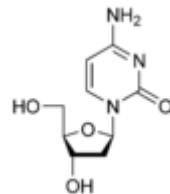
Nitrogenous base	Nucleoside	Deoxynucleoside
		
Adenine	Adenosine A	Deoxyadenosine dA
		
Guanine	Guanosine G	Deoxyguanosine dG
		
Thymine	5-Methyluridine m^5U	Thymidine dT
		
Uracil	Uridine U	Deoxyuridine dU



Cytosine



Cytidine
C



Deoxycytidine
dC

Nucleosides are glycosylamines consisting of a nucleobase (often referred to as simply base) bound to a ribose or deoxyribose sugar via a beta-glycosidic linkage. Examples of nucleosides include cytidine, uridine, adenosine, guanosine, thymidine and inosine. Nucleosides can be phosphorylated by specific kinases in the cell on the sugar's primary alcohol group (-CH₂-OH), producing nucleotides, which are the molecular building-blocks of DNA and RNA.

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

1. Adenosine is a nucleoside composed of a molecule of adenine attached to a ribose sugar molecule (ribofuranose) moiety via a β -N9-glycosidic bond.
2. Cytidine is a nucleoside molecule that is formed when cytosine is attached to a ribose ring (also known as a ribofuranose) via a β -N1-glycosidic bond. Cytidine is a component of RNA.
3. Guanosine is a purine nucleoside comprising guanine attached to a ribose (ribofuranose) ring via a β -N9-glycosidic bond. Guanosine can be phosphorylated to become guanosine monophosphate (GMP), cyclic guanosine monophosphate (cGMP), guanosine diphosphate (GDP), and guanosine triphosphate (GTP).

4. Thymidine (more precisely called deoxythymidine; can also be labelled deoxyribosylthymine, and thymine deoxyriboside) is a chemical compound, more precisely a pyrimidine deoxynucleoside. Deoxythymidine is the DNA nucleoside T, which pairs with deoxyadenosine (A) in double-stranded DNA.

If cytosine is attached to a deoxyribose ring, it is known as a deoxycytidine^[9]

Nucleotide

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.

A deoxyribonucleotide is the monomer, or single unit, of DNA, or deoxyribonucleic acid. Each deoxyribonucleotide comprises three parts: a nitrogenous base, a deoxyribose sugar, and one or more phosphate groups. The nitrogenous base is always bonded to the 1' carbon of the deoxyribose, which is distinguished from ribose by the presence of a proton on the 2' carbon rather than an -OH group. The phosphate groups bind to the 5' carbon of the sugar. When

deoxyribonucleotides polymerize to form DNA, the phosphate group from one nucleotide will bond to the 3' carbon on another nucleotide, forming a phosphodiester bond via dehydration synthesis. New nucleotides are always added to the 3' carbon of the last nucleotide, so synthesis always proceeds from 5' to 3'.^[10]

Phosphodiester bond

A phosphodiester bond is a group of strong covalent bonds between a phosphate group and two 5-carbon ring carbohydrates (pentoses) over two ester bonds. Phosphodiester bonds are central to most life on Earth, as they make up the backbone of the strands of DNA. In DNA and RNA, the phosphodiester bond is the linkage between the 3' carbon atom of one sugar molecule and the 5' carbon of another, deoxyribose in DNA and ribose in RNA. The phosphate groups in the phosphodiester bond are negatively-charged. Because the phosphate groups have a pKa near 0, they are negatively-charged at pH 7. This repulsion forces the phosphates to take opposite sides of the DNA strands and is neutralized by proteins (histones), metal ions such as magnesium, and polyamines. In order for the phosphodiester bond to be formed and the nucleotides to be joined, the tri-phosphate or di-phosphate forms of the nucleotide building blocks are broken apart to give off energy required to drive the enzyme-catalyzed reaction. When a single phosphate or two phosphates known as pyrophosphates break away and catalyze the reaction, the phosphodiester bond is formed. Hydrolysis of phosphodiester bonds can be catalyzed by the action of phosphodiesterases which play an important role in repairing DNA sequences. In biological systems, the phosphodiester bond between two ribonucleotides can be broken by alkaline hydrolysis because of the free 2' hydroxyl group.^[11]

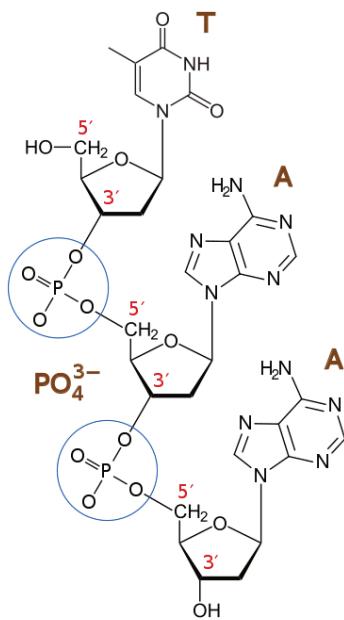
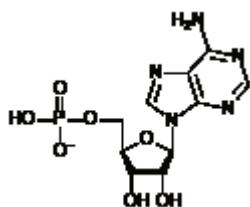
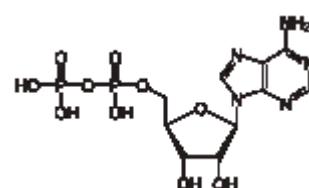


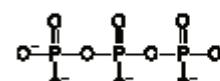
Diagram of phosphodiester bonds (PO_4^{3-}) between nucleotides. Which presents Thymine (U) and two molecules of Adenine (A).



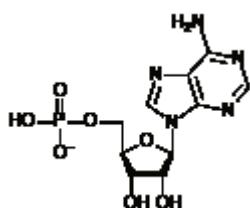
Adenosine
monophosphate
AMP



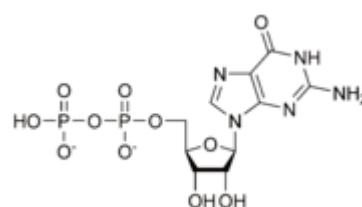
Adenosine diphosphate
ADP



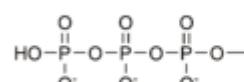
Adenosine triphosphate
ATP



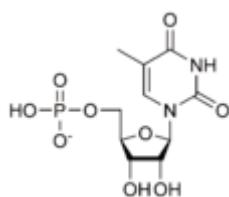
Guanosine
monophosphate
GMP



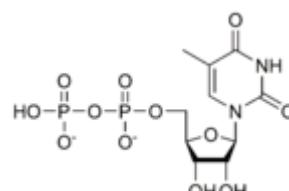
Guanosine diphosphate
GDP



Guanosine triphosphate
GTP



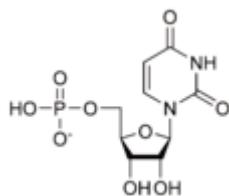
Ribothymidine
monophosphate
rTMP



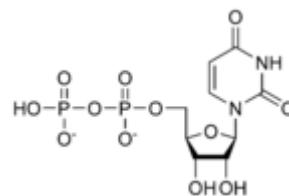
Ribothymidine diphosphate
rTDP



Ribothymidine triphosphate
rTTP



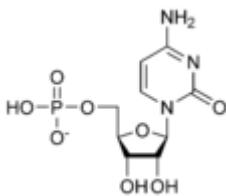
Uridine monophosphate
UMP



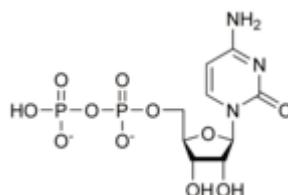
Uridine diphosphate
UDP



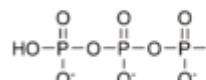
Uridine triphosphate
UTP



Cytidine
monophosphate
CMP



Cytidine diphosphate
CDP



Cytidine triphosphate
CTP

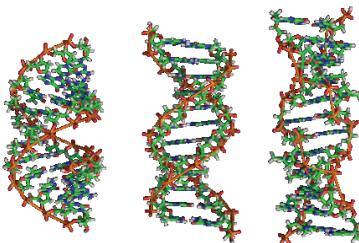
Forms of DNA

A-DNA: A-DNA is one of the many possible double helical structures of DNA. A-DNA is thought to be one of three biologically active double helical structures along with B- and Z-DNA. It is a right-handed double helix fairly similar to the more common and well-known B-DNA form, but with a shorter more compact helical structure. It appears likely that it occurs only in dehydrated samples of DNA, such as those used in crystallographic experiments, and possibly is also assumed by DNA-RNA hybrid helices and by regions of double-stranded RNA.^[12]

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

Z-DNA: Z-DNA is one of the many possible double helical structures of DNA. It is a left-handed double helical structure in which the double helix winds to the left in a zig-zag pattern (instead of to the right, like the more common B-DNA form). Z-DNA is

thought to be one of three biologically active double helical structures along with A- and B-DNA. Z-DNA is quite different from the right-handed forms. In fact, Z-DNA is often compared against B-DNA in order to illustrate the major differences. The Z-DNA helix is left-handed and has a structure that repeats every 2 base pairs. The major and minor grooves, unlike A- and B-DNA, show little difference in width. Formation of this structure is generally unfavourable, although certain conditions can promote it; such as alternating purine-pyrimidine sequence (especially poly(dGC)2), negative DNA supercoiling or high salt and some cations (all at physiological temperature, 37 °C, and pH 7.3-7.4). Z-DNA can form a junction with B-DNA (called a “B-to-Z junction box”) in a structure which involves the extrusion of a base pair. The Z-DNA conformation has been difficult to study because it does not exist as a stable feature of the double helix. Instead, it is a transient structure that is occasionally induced by biological activity and then quickly disappears.^[13]



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

Difference between three major forms of DNA

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

bp-Base pair, nm-nano meter

Noncoding genomic DNA

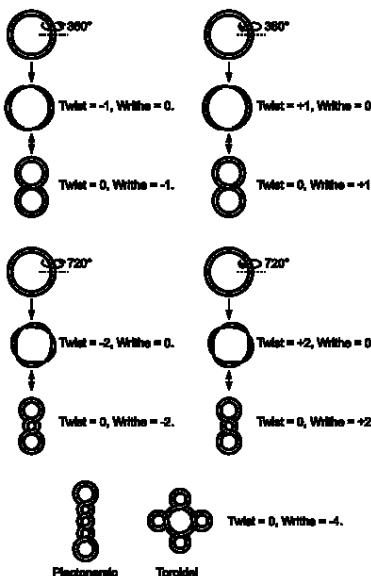
In molecular biology, noncoding DNA describes components of an organism's DNA sequences that do not encode for protein sequences.

Pseudogenes Pseudogenes are DNA sequences, related to known genes, that have lost their protein-coding ability or are otherwise no longer expressed in the cell. Pseudogenes arise from retrotransposition or genomic duplication of functional genes, and become "genomic fossils" that are nonfunctional due to mutations that prevent the transcription of the gene, such as within the gene promoter region, or fatally alter the translation of the gene, such

as premature stop codons or frameshifts. Pseudogenes resulting from the retrotransposition of an RNA intermediate are known as processed pseudogenes; pseudogenes that arise from the genomic remains of duplicated genes or residues of inactivated genes are nonprocessed pseudogenes. While Dollo's Law suggests that the loss of function in pseudogenes is likely permanent, silenced genes may actually retain function for several million years and can be "reactivated" into protein-coding sequences and a substantial number of pseudogenes are actively transcribed. Because pseudogenes are presumed to evolve without evolutionary constraint, they can serve as a useful model of the type and frequencies of various spontaneous genetic mutations.^[14]

Coiling of DNA

DNA supercoiling is important for DNA packaging within all cells. Because the length of DNA can be thousands of times that of a cell, packaging this genetic material into the cell or nucleus (in eukaryotes) is a difficult feat. Supercoiling of DNA reduces the space and allows for a lot more DNA to be packaged. In prokaryotes, plectonemic supercoils are predominant, because of the circular chromosome and relatively small amount of genetic material. In eukaryotes, DNA supercoiling exists on many levels of both plectonemic and solenoidal supercoils, with the solenoidal supercoiling proving most effective in compacting the DNA. Solenoidal supercoiling is achieved with histones to form a 10 nm fiber. This fiber is further coiled into a 30 nm fiber, and further coiled upon itself numerous times more. DNA packaging is greatly increased during nuclear division events such as mitosis or meiosis, where DNA must be compacted and segregated to daughter cells. Condensins and cohesins are Structural Maintenance of Chromosome proteins that aid in the condensation of sister chromatids and the linkage of the centromere in sister chromatids. These SMC proteins induce positive supercoils. Supercoiling is also required for DNA/RNA synthesis. Because DNA must be unwound for DNA/RNA polymerase action, supercoils will result. The region



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

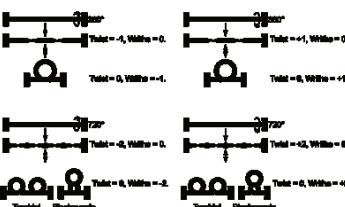
ahead of the polymerase complex will be unwound; this stress is compensated with positive supercoils ahead of the complex. Behind the complex, DNA is rewound and there will be compensatory negative supercoils. It is important to note that topoisomerases such as DNA gyrase (Type II Topoisomerase) play a role in relieving some of the stress during DNA/RNA synthesis.^[15]

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

$$\Delta Lk = \frac{\text{bp}}{10.4}$$

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

$$\Delta Lk = TW - Wr$$



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

$$\Delta Lk = Lk - Lk_{\text{relaxed}}$$

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

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

A standard expression independent of the molecule size is the

“specific linking difference” or “superhelical density” denoted σ . σ represents the number of turns added or removed relative to the total number of turns in the relaxed molecule/plasmid, indicating the level of supercoiling.

$$\sigma = \Delta \frac{Lk}{Lk_0}$$

The Gibbs free energy associated with the coiling is given by the equation below^[16]

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

The linking number is a numerical invariant that describes the linking of two closed curves in three-dimensional space. Intuitively, the linking number represents the number of times that each curve winds around the other. The linking number is always an integer, but may be positive or negative depending on the orientation of the two curves. Since the linking number L of supercoiled DNA is the number of times the two strands are intertwined (and both strands remain covalently intact), L cannot change. The reference state (or parameter) L_0 of a circular DNA duplex is its relaxed state. In this state, its writhe $W = 0$. Since $L = T W$, in a relaxed state $T = L$. Thus, if we have a 400 bp relaxed circular DNA duplex, $L \sim 40$ (assuming ~10 bp per turn in B-DNA). Then $T \sim 40$.

- Positively supercoiling:

$$T = 0, W = 0, \text{ then } L = 0$$

$$T = 3, W = 0, \text{ then } L = 3$$

$$T = 2, W = 1, \text{ then } L = 3$$

- Negatively supercoiling:

$$T = 0, W = 0, \text{ then } L = 0$$

$$T = -3, W = 0, \text{ then } L = -3$$

$$T = -2, W = -1, \text{ then } L = -3$$

Negative supercoils favor local unwinding of the DNA, allowing processes such as transcription, DNA replication, and recombination. Negative supercoiling is also thought to favour the

transition between B-DNA and Z-DNA, and moderate the interactions of DNA binding proteins involved in gene regulation.^[17]

Histones: The DNA binding protein

Histones were discovered in 1884 by Albrecht Kossel. The word “histone” dates from the late 19th century and is from the German “Histon”, of uncertain origin: perhaps from Greek *histanai* or from *histos*. Until the early 1990s, histones were dismissed by most as inert packing material for eukaryotic nuclear DNA, based in part on the “ball and stick” models of Mark Ptashne and others who believed transcription was activated by protein-DNA and protein-protein interactions on largely naked DNA templates, as is the case in bacteria. During the 1980s, work by Michael Grunstein^[18] demonstrated that eukaryotic histones repress gene transcription, and that the function of transcriptional activators is to overcome this repression. We now know that histones play both positive and negative roles in gene expression, forming the basis of the histone code.

The discovery of the H5 histone appears to date back to 1970's,^{[19][20]} and in classification it has been grouped with The nucleosome core is formed of two H2A-H2B dimers and a H3-H4 tetramer, forming two nearly symmetrical halves by tertiary structure (C2 symmetry; one macromolecule is the mirror image of the other). The H2A-H2B dimers and H3-H4 tetramer also show pseudodyad symmetry. The 4 ‘core’ histones (H2A, H2B, H3 and H4) are relatively similar in structure and are highly conserved through evolution, all featuring a ‘helix turn helix turn helix’ motif (which allows the easy dimerisation). They also share the feature of long ‘tails’ on one end of the amino acid structure – this being the location of post-translational modification (see below).

It has been proposed that histone proteins are evolutionarily related to the helical part of the extended AAA ATPase domain, the C-domain, and to the N-terminal substrate recognition domain of Clp/Hsp100 proteins. Despite the differences in their topology, these three folds share a homologous helix-strand-helix (HSH) motif.

Using an electron paramagnetic resonance spin-labeling technique, British

researchers measured the distances between the spools around which eukaryotic cells wind their DNA. They determined the spacings range from 59 to 70 Å. In all, histones make five types of interactions with DNA:

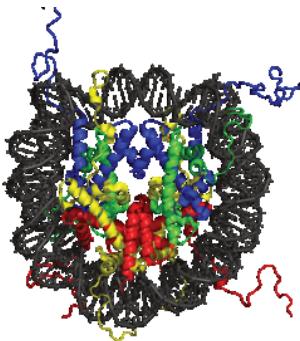
Helix-dipoles from alpha-helices in H2B, H3, and H4 cause a net positive charge to accumulate at the point of interaction with negatively charged phosphate groups on DNA. Hydrogen bonds between the DNA backbone and the amide group on the main chain of histone proteins

Nonpolar interactions between the histone and deoxyribose sugars on DNA

Salt bridges and hydrogen bonds between side chains of basic amino acids (especially lysine and arginine) and phosphate oxygens on DNA

Non-specific minor groove insertions of the H3 and H2B N-terminal tails into two minor grooves each on the DNA molecule

The highly basic nature of histones, aside from facilitating DNA-histone interactions, contributes to the water solubility of histones. Histones are subject to post translational modification by enzymes primarily on their N-terminal tails, but also in their globular



The crystal structure of the nucleosome core particle consisting of H2A, H2B, H3 and H4 and DNA. The view is from the top through the superhelical axis.

domains. Such modifications include methylation, citrullination, acetylation, phosphorylation, SUMOylation, ubiquitination, and ADP-ribosylation. This affects their function of gene regulation. In general, genes that are active have less bound histone, while inactive genes are highly associated with histones during interphase. It also appears that the structure of histones has been evolutionarily conserved, as any deleterious mutations would be severely maladaptive.

Histone DNA interaction

The core histone proteins contain a characteristic structural motif termed the “histone fold” which consists of three alpha-helices (α 1-3) separated by two loops (L1-2). In solution the histones form H2A-H2B heterodimers and H3-H4 heterotetramers. Histones dimerise about their long α 2 helices in an anti-parallel orientation, and in the case of H3 and H4, two such dimers form a 4-helix bundle stabilised by extensive H3-H3' interaction. The H2A/H2B dimer binds onto the H3/H4 tetramer due to interactions between H4 and H2B which include the formation of a hydrophobic cluster. The histone octamer is formed by a central H3/H4 tetramer sandwiched between two H2A/H2B dimers. Due to the highly basic charge of all four core histones, the histone octamer is only stable in the presence of DNA or very high salt concentrations.

Nucleosomes form the fundamental repeating units of eukaryotic chromatin, which is used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 μm diameter). Nucleosomes are folded through a series of successively higher order structures to eventually form a chromosome; this both compacts DNA and creates an added layer of regulatory control which ensures correct gene expression. Nucleosomes are thought to carry epigenetically

inherited information in the form of covalent modifications of their core histones. The nucleosome hypothesis was proposed by Don and Ada Olins in 1974 and Roger Kornberg.

The nucleosome core particle) consists of about 146 bp of DNA wrapped in 1.67 left-handed superhelical turns around the histone octamer, consisting of 2 copies each of the core histones H2A, H2B, H3, and H4. Adjacent nucleosomes are joined by a stretch of free DNA termed “linker DNA” (which varies from 10 – 80 bp in length depending on species and tissue type.

DNA-binding domains

One or more DNA-binding domains are often part of a larger protein consisting of additional domains with differing function. The additional domains often regulate the activity of the DNA-binding domain. The function of DNA binding is either structural or involving transcription regulation, with the two roles sometimes overlapping. DNA-binding domains with functions involving DNA structure have biological roles in the replication, repair, storage, and modification of DNA, such as methylation. Many proteins involved in the regulation of gene expression contain DNA-binding domains. For example, proteins that regulate transcription by binding DNA are called transcription factors. The final output of most cellular signaling cascades is gene regulation. The DBD interacts with the nucleotides of DNA in a DNA sequence-specific or non-sequence-specific manner, but even non-sequence-specific recognition involves some sort of molecular complementarity between protein and DNA. DNA recognition by the DBD can occur at the major or minor groove of DNA, or at the sugar-



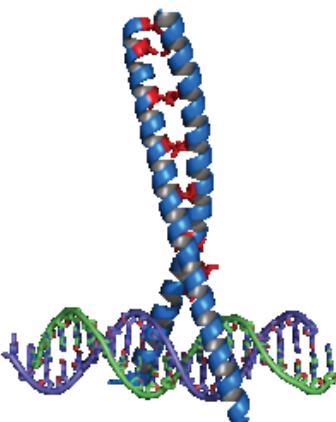
The λ repressor of bacteriophage lambda employs a helix-turn-helix (left; green) to bind DNA (right; blue and red).

phosphate DNA backbone (see the structure of DNA). Each specific type of DNA recognition is tailored to the protein's function. For example, the DNA-cutting enzyme DNase I cuts DNA almost randomly and so must bind to DNA in a non-sequence-specific manner. But, even so, DNase I recognizes a certain 3-D DNA structure, yielding a somewhat specific DNA cleavage pattern that can be useful for studying DNA recognition by a technique called DNA footprinting. Many DNA-binding domains must recognize specific DNA sequences, such as DBDs of transcription factors that activate specific genes, or those of enzymes that modify DNA at specific sites, like restriction enzymes and telomerase. The hydrogen bonding pattern in the DNA major groove is less degenerate than that of the DNA minor groove, providing a more attractive site for sequence-specific DNA recognition. The specificity of DNA-binding proteins can be studied using many biochemical and biophysical techniques, such as gel electrophoresis, analytical ultracentrifugation, calorimetry, DNA mutation, protein structure mutation or modification, nuclear magnetic resonance, x-ray crystallography, surface plasmon resonance, electron paramagnetic resonance, cross-linking and Microscale Thermophoresis (MST).^[21]

Types of DNA-binding domains

Helix-turn-helix

Originally discovered in bacteria, the helix-turn-helix motif is commonly found in repressor proteins and is about 20 amino acids long. In eukaryotes, the homeodomain comprises 2 helices, one of which recognizes the DNA (aka recognition helix). They are common in proteins that regulate developmental processes (PROSITE HTH).[22]



Leucine Zipper (blue) bound to DNA.
The leucine residues that represent the 'teeth' of the zipper are colored red

Zinc finger

Crystallographic structure (PDB 1R4O) of a dimer of the zinc finger containing DBD of the glucocorticoid receptor (top) bound to DNA (bottom). Zinc atoms are represented by grey spheres and the coordinating cysteine sidechains are depicted as sticks. The zinc finger This domain is generally between 23 and 28 amino acids long and is stabilized by coordinating Zinc ions with regularly spaced zinc-coordinating residues (either histidines or cysteines). The most common class of zinc finger (Cys2His2) coordinates a single zinc ion and consists of a recognition helix and a 2-strand beta-sheet. In transcription factors these domains are often found in arrays (usually separated by short linker sequences) and adjacent fingers are spaced at 3 basepair intervals when bound to DNA.



Crick and Watson DNA model built in 1953, was largely from its original pieces in 1973 and donated to the National Science Museum in London.

Fold Group	Representative structure	Ligand placement
Cys ₂ His ₂		Two ligands from a knuckle and two more from the C terminus of a helix.
Gag knuckle		Two ligands from a knuckle and two more from a short helix or loop.
Treble clef		Two ligands from a knuckle and two more from the N terminus of a helix.
Zinc ribbon		Two ligands each from two knuckles.
Zn ₂ /Cys ₆		Two ligands from the N terminus of a helix and two more from a loop.

TAZ2
domain
like

Two
ligands
from the
termini of
two
helices.

Leucine zipper

The basic leucine zipper (bZIP) domain contains an alpha helix with a leucine at every 7th amino acid. If two such helices find one another, the leucines can interact as the teeth in a zipper, allowing dimerization of two proteins. When binding to the DNA, basic amino acid residues bind to the sugar-phosphate backbone while the helices sit in the major grooves. It regulates gene expression. The bZip family of transcription factors consist of a basic region that interacts with the major groove of a DNA molecule through hydrogen bonding, and a hydrophobic leucine zipper region that is responsible for dimerization.

Winged helix

Consisting of about 110 amino acids, the winged helix (WH) domain has four helices and a two-strand beta-sheet.

Winged helix turn helix

The winged helix turn helix domain (wHTH) SCOP 46785 is typically 85–90 amino acids long. It is formed by a 3-helical bundle and a 4-strand beta-sheet (wing).

Helix-loop-helix

The Helix-loop-helix domain is found in some transcription factors and is characterized by two α helices connected by a loop. One helix is typically smaller and due to the flexibility of the loop, allows dimerization by folding and packing against another helix. The larger helix typically contains the DNA-binding regions.

HMG-box

HMG-box domains are found in high mobility group proteins which are involved in a variety of DNA-dependent processes like replication and transcription. The domain consists of three alpha helices separated by loops.

DNA sequencing

RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), between 1972 and 1976. Prior to the development of rapid DNA sequencing methods in the early 1970s by Frederick Sanger at the University of Cambridge, in England and Walter Gilbert and Allan Maxam at Harvard, a number of laborious methods were used. For instance, in 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis. The chain-termination method developed by Sanger and coworkers in 1975 soon became the method of choice, owing to its relative ease and reliability.^[23]

Maxam and Gilbert method

In 1976–1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method (see below), Maxam–Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-

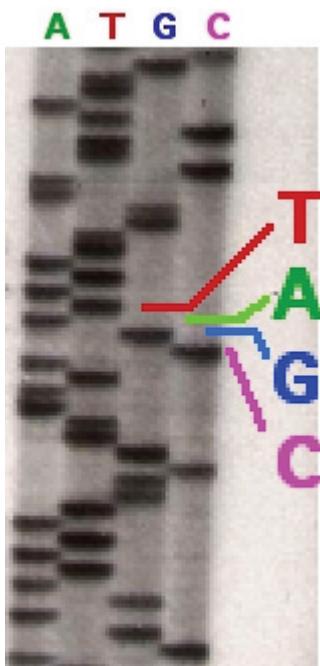
up. The method requires radioactive labeling at one 5' end of the DNA (typically by a kinase reaction using gamma-³²P ATP) and purification of the DNA fragment to be sequenced. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A G, C, C T). For example, the purines (A G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C T) are methylated using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction. The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule. The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred. Also sometimes known as "chemical sequencing", this method led to the Methylation Interference Assay used to map DNA-binding sites for DNA-binding proteins.^[24]

Dideoxynucleotide Chain-termination methods

Because the chain-terminator method (or Sanger method after its developer Frederick Sanger) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice. The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

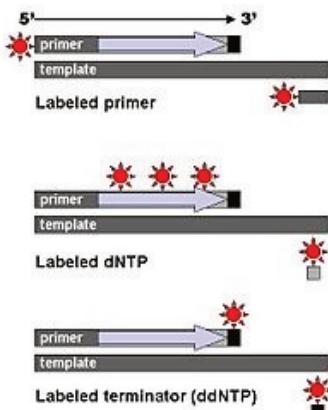
The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotidetriphosphates (dNTPs), and modified nucleotides (dideoxyNTPs) that terminate DNA strand elongation. These ddNTPs will also be radioactively or fluorescently labelled for detection in automated sequencing machines. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-hydroxyl (OH) group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

The newly synthesized and labelled DNA fragments are heat



Part of a radioactively labelled sequencing gel

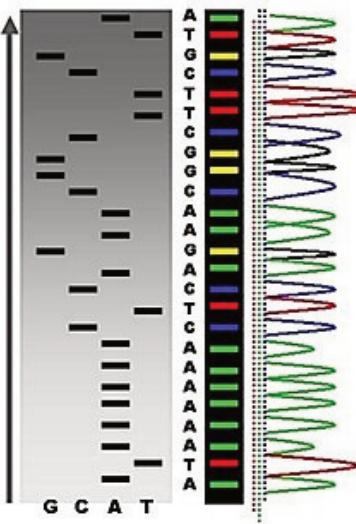
denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.^[25]



DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP. (click to expand)

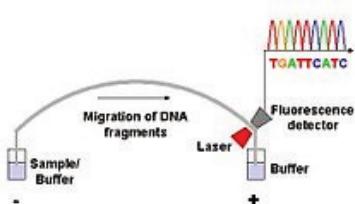
Technical variations of chain-termination sequencing include tagging with nucleotides containing phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development by Leroy Hood and coworkers [26][27] of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

Dye-terminator sequencing



Capillary electrophoresis (click to expand)

chain terminators is labelled with fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide

incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis (see figure to the left).

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating “dye blobs”. The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

Challenges

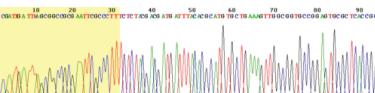
Common challenges of DNA sequencing include poor quality in the first 15–40 bases of the sequence and deteriorating quality of sequencing traces after 700–900 bases. Base calling software typically gives an estimate of quality to aid in quality trimming.^{[28][29]}

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and emerging sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.^{[30][31]}

Current methods can directly sequence only relatively short (300–1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide. In all cases the use of a primer with a free 5' end is essential.

Automation and sample preparation

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch (run) in up to 24 runs a day. DNA



View of the start of an example dye-terminator read

sequencers carry out capillary electrophoresis for size separation, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms. Sequencing reactions by thermocycling, cleanup and re-suspension in a buffer solution before loading onto the sequencer are performed separately. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (generally located at the ends of the sequence). The accuracy of such algorithms is below visual examination by a human operator, but sufficient for automated processing of large sequence data sets.

Polymerase chain reaction

PCR

PCR is used to amplify a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of up to ~10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size. A basic PCR set up requires several components and reagents. These components include:

DNA template that contains the DNA region (target) to be amplified.

Two primers that are complementary to the 3' (three prime) ends of each of the sense and anti-sense strand of the DNA target. Taq polymerase or another DNA polymerase with a temperature optimum at around 70 °C. Deoxynucleotide triphosphates (dNTPs), the building-blocks from which the DNA polymerase synthesizes a new DNA strand. Buffer solution, providing a suitable chemical environment for optimum activity and stability of the DNA polymerase.

Divalent cations, magnesium or manganese ions; generally Mg²⁺ is used, but Mn²⁺ can be utilized for PCR-mediated DNA mutagenesis, as higher Mn²⁺ concentration

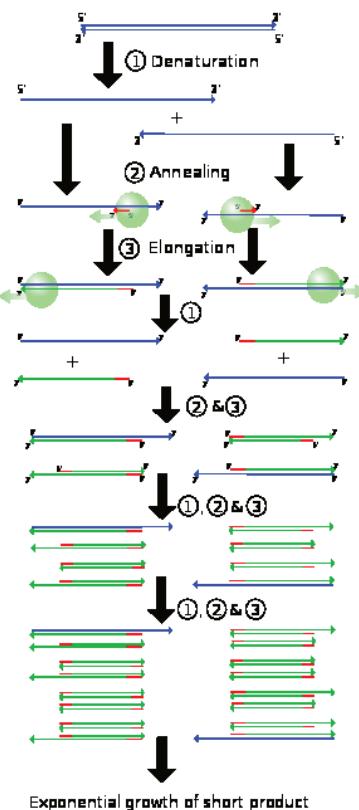


Figure 1: Schematic drawing of the PCR cycle. (1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C.

Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give shorter DNA products (green lines), which themselves are used as templates as PCR progresses.

increases the error rate during DNA synthesis. Monovalent cation potassium ions. The PCR is commonly carried out in a reaction volume of 10–200 µl in small reaction tubes (0.2–0.5 ml volumes) in a thermal cycler. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction (see below). Many modern thermal cyclers make use of the Peltier effect, which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled reaction tubes permit favorable thermal conductivity to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermocyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.^[32]

Procedure

Figure 1: Schematic drawing of the PCR cycle. (1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C. Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give shorter DNA products (green lines), which themselves are used as templates as PCR progresses. Typically, PCR consists of a series of 20–40 repeated temperature changes, called cycles, with each cycle commonly consisting of 2–3 discrete temperature steps, usually three. The cycling is often preceded by a single temperature step (called hold) at a high temperature (>90 °C), and followed by one hold at the end for final product extension or brief storage. The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis, the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (T_m) of the primers. Initialization step: This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes. It is only required for DNA polymerases that require heat activation by hot-start PCR. Denaturation step: This step is the first regular cycling

event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules. Annealing step: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically the annealing temperature is about 3–5 degrees Celsius below the Tm of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis. Extension/elongation step: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to be amplified. As a rule-of-thumb, at its optimum temperature, the DNA polymerase will polymerize a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment. Final elongation: This single step is occasionally performed at a temperature of 70–74 °C for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended. Final hold: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.

To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplicon or amplicon), agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by

comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products.

Facts to be remembered

DNA Polymerases are enzymes that synthesize polynucleotide chains from nucleoside triphosphates and make the DNA. In 1865 Gregor Mendel's paper, Experiments on Plant Hybridization

In 1869, DNA was first isolated by the Swiss physician **Friedrich Miescher** who discovered a microscopic substance in the pus of discarded surgical bandages.

From 1880-1890 Walther Flemming, Eduard Strasburger, and Edouard van Beneden elucidate chromosome distribution during cell division

In 1889 Hugo de Vries postulates that “inheritance of specific traits in organisms comes in particles”, naming such particles “(pan)genes”

In 1903 Walter Sutton hypothesizes that chromosomes, which segregate in a Mendelian fashion, are hereditary units

In 1905 William Bateson coins the term “genetics” in a letter to Adam Sedgwick and at a meeting in 1906

In 1908 Hardy-Weinberg law derived.

In 1910 Thomas Hunt Morgan shows that genes reside on chromosomes

In 1913 Alfred Sturtevant makes the first genetic map of a chromosome

In 1913 Gene maps show chromosomes containing linear arranged genes

In 1918 Ronald Fisher publishes “The Correlation Between Relatives on the Supposition of Mendelian Inheritance” the modern synthesis of genetics and evolutionary biology starts. See population genetics.

In 1928 Frederick Griffith discovers that hereditary material from dead bacteria can be incorporated into live bacteria (see Griffith's experiment)

In 1931 Crossing over is identified as the cause of recombination

In 1933 Jean Brachet is able to show that DNA is found in chromosomes and that RNA is present in the cytoplasm of all cells.

In 1937 William Astbury produced the first X-ray diffraction patterns that showed that DNA had a regular structure.

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.

In 1952, **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 double-helix model of DNA structure.

Purines are found in high concentration in meat and meat products, especially internal organs such as liver and kidney.

Examples of high-purine sources include: sweetbreads, anchovies, sardines, liver, beef kidneys, brains, meat extracts (e.g., Oxo, Bovril), herring, mackerel, scallops, game meats, beer (from the yeast) and gravy.

bp = base pair(s) One bp corresponds to circa 3.4 Å of length along the strand

kb (= kbp) = kilo base pairs = 1,000 bp

Mb = mega base pairs = 1,000,000 bp

Analysis of DNA topology uses three values:

L = linking number – the number of times one DNA strand wraps around the other. It is an integer for a closed loop and constant for a closed topological domain.

T = twist – total number of turns in the double stranded DNA helix. This will normally tend to approach the number of turns that a topologically open double stranded DNA helix makes free in solution: number of bases/10.5, assuming there are no intercalating

agents (e.g., chloroquine) or other elements modifying the stiffness of the DNA.

W = writhe – number of turns of the double stranded DNA helix around the superhelical axis

$$L = TW \text{ and } \Delta L = \Delta T \Delta W$$

Any change of T in a closed topological domain must be balanced by a change in W , and vice versa. This results in higher order structure of DNA. A circular DNA molecule with a writhe of 0 will be circular. If the twist of this molecule is subsequently increased or decreased by supercoiling then the writhe will be appropriately altered, making the molecule undergo plectonemic or toroidal superhelical coiling. When the ends of a piece of double stranded helical DNA are joined so that it forms a circle the strands are topologically knotted. This means the single strands cannot be separated any process that does not involve breaking a strand (such as heating). The task of un-knotting topologically linked strands of DNA falls to enzymes known as topoisomerases. These enzymes are dedicated to un-knotting circular DNA by cleaving one or both strands so that another double or single stranded segment can pass through. This un-knotting is required for the replication of circular DNA and various types of recombination in linear DNA which have similar topological constraints.

$$Gb = \text{giga base pairs} = 1,000,000,000 \text{ bp.}$$

1972 Development of recombinant DNA technology, which permits isolation of defined fragments of DNA; prior to this, the only accessible samples for sequencing were from bacteriophage or virus DNA.

1977 The first complete DNA genome to be sequenced is that of bacteriophage ϕ X174. 1977 Allan Maxam and Walter Gilbert publish “DNA sequencing by chemical degradation”. Frederick Sanger, independently, publishes “DNA sequencing with chain-terminating inhibitors”. 1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.

1986 Leroy E. Hood's laboratory at the California Institute of

Technology and Smith announce the first semi-automated DNA sequencing machine.

1987 Applied Biosystems markets first automated sequencing machine, the model ABI 370.

1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (at US\$0.75/base).

1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.

1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) publish the first complete genome of a free-living organism, the bacterium *Haemophilus influenzae*. The circular chromosome contains 1,830,137 bases and its publication in the journal *Science* marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

1996 Pål Nyrén and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm publish their method of pyrosequencing. 1998 Phil Green and Brent Ewing of the University of Washington publish "phred" for sequencer data analysis.

2001 A draft sequence of the human genome is published.

2004 454 Life Sciences markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of a new generation of sequencing technologies, after MPSS

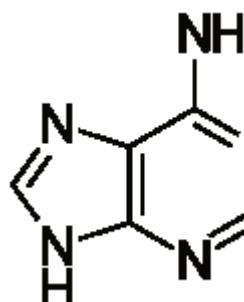
List of bases found in DNA and RNA

Name	3-D structure	Abbreviation	Structural formula
Cytosine		C	
Thymine		T	
Uracil		U	

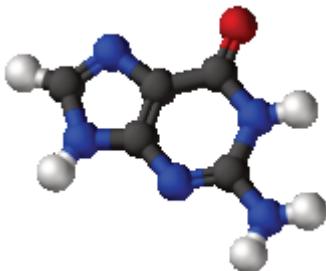
Adenine



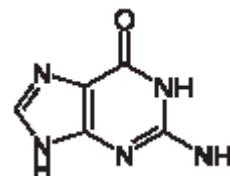
A



Guanine



C



References

1. ↑ DNA
2. ↑ Griffith experiment
3. ↑ Hershey–Chase experiment
4. ↑ Hershey, A.D. and Chase, M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol.* 36:39–56.
5. ↑ Avery–MacLeod–McCarty experiment
6. ↑ Base pair
7. ↑ Pyrimidine
8. ↑ Cytosine
9. ↑ Nucleoside
10. ↑ Nucleotide

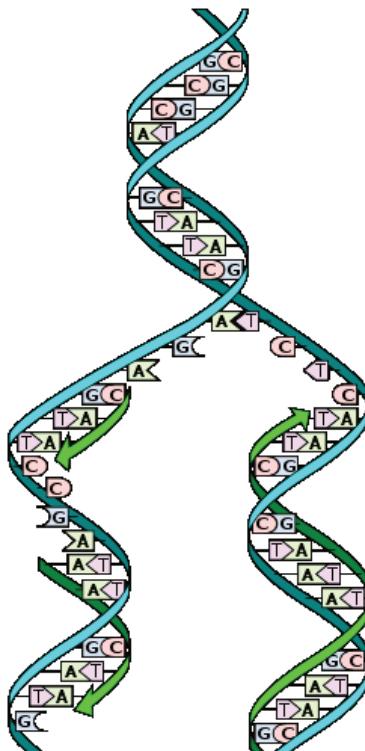
11. ↑ Phosphodiester bond
12. ↑ A-DNA
13. ↑ <http://en.wikipedia.org/wiki/Z-DNA>
14. ↑ Noncoding DNA
15. ↑ DNA supercoil
16. ↑ Vologodskii AV, Lukashin AV, Anshelevich VV, et al. (1979). "Fluctuations in superhelical DNA". *Nucleic Acids Res* **6**: 967–682. doi:10.1093/nar/6.3.967.
17. ↑ H. S. Chawla (2002). *Introduction to Plant Biotechnology*. Science Publishers. ISBN1578082285.
18. ↑ Kayne PS, Kim UJ, Han M, Mullen JR, Yoshizaki F, Grunstein M. Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. *Cell*. 1988 Oct 7;55(1):27-39. PMID 3048701
19. ↑ Crane-Robinson C, Dancy SE, Bradbury EM, Garel A, Kovacs AM, Champagne M, Daune M (August 1976). "Structural studies of chicken erythrocyte histone H5". *Eur. J. Biochem.* **67** (2): 379–88. doi:10.1111/j.1432-1033.1976.tb10702.x. PMID964248.
20. ↑ Aviles FJ, Chapman GE, Kneale GG, Crane-Robinson C, Bradbury EM (August 1978). "The conformation of histone H5. Isolation and characterisation of the globular segment". *Eur. J. Biochem.* **88** (2): 363–71. doi:10.1111/j.1432-1033.1978.tb12457.x. PMID689022.
21. ↑ http://en.wikipedia.org/wiki/DNA-binding_domain
22. ↑ http://en.wikipedia.org/wiki/DNA-binding_domain
23. ↑ http://en.wikipedia.org/wiki/DNA_sequencing
24. ↑ http://en.wikipedia.org/wiki/DNA_sequencing
25. ↑ DNA sequencing
26. ↑ Smith LM, Sanders JZ, Kaiser RJ, et al (1986). "Fluorescence detection in automated DNA sequence analysis". *Nature* **321** (6071): 674–9. doi:10.1038/321674a0. PMID3713851.
"We have developed a method for the partial automation of DNA sequence analysis. Fluorescence detection of the DNA fragments is accomplished by means of a fluorophore covalently attached to the oligonucleotide primer used in

enzymatic DNA sequence analysis. A different coloured fluorophore is used for each of the reactions specific for the bases A, C, G and T. The reaction mixtures are combined and co-electrophoresed down a single polyacrylamide gel tube, the separated fluorescent bands of DNA are detected near the bottom of the tube, and the sequence information is acquired directly by computer.”.

27. ↑ Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE (April 1985). “The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis”. *Nucleic Acids Res.* **13** (7): 2399–412. doi:10.1093/nar/13.7.2399. PMID4000959. PMC341163. <http://nar.oxfordjournals.org/cgi/pmidlookup?view=long&pmid=4000959>.
28. ↑ “Phred – Quality Base Calling”. <http://www.phrap.com/phred/>. Retrieved 2011-02-24.
29. ↑ “Base-calling for next-generation sequencing platforms – Brief Bioinform”. <http://bib.oxfordjournals.org/content/early/2011/01/18/bib.bbq077.full>. Retrieved 2011-02-24.
30. ↑ Murphy, K.; Berg, K.; Eshleman, J. (2005). “Sequencing of genomic DNA by combined amplification and cycle sequencing reaction”. *Clinical chemistry* 51 (1): 35–39.
31. ↑ Sengupta, D.; Cookson, B. (2010). “SeqSharp: A general approach for improving cycle-sequencing that facilitates a robust one-step combined amplification and sequencing method”. *The Journal of molecular diagnostics : JMD* 12 (3): 272–277.
32. ↑ Polymerase chain reaction

I3.

As we know Cell division is essential for an organism to grow, but, when a cell divides, it must replicate the DNA (DNA replication take place during S phase) 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.



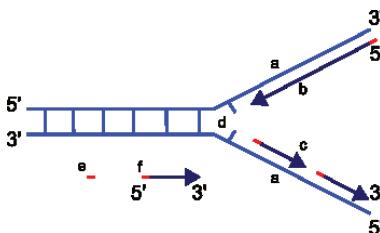
DNA replication. DNA is unwound and nucleotides are matched to make two new strands.

Contents

- 1 Replication
 - 1.1 Leading strand
 - 1.2 Lagging strand
 - 1.2.1 Okazaki fragment
 - 1.3 Rate of replication
 - 1.4 Classification of DNA polymerase
- 2 Replication is semiconservative
- 3 Replication in prokaryote
 - 3.1 Primase
 - 3.2 Primosome
 - 3.3 Elongation of DNA strand
 - 3.4 Termination
- 4 DNA polymerase in prokaryote
 - 4.1 DNA Polymerase I or Pol I
 - 4.1.1 Klenow fragment
- 5 Helicases and Topoisomerase
- 6 Replication in Eukaryote
 - 6.1 Eukaryotic DNA polymerase
 - 6.2 DNA Replication occurs during the S phase
- 7 Replication in mitochondria
- 8 C_{ot} values
- 9 DNA repair
 - 9.1 Types of DNA damage
 - 9.2 Sources of damage
 - 9.3 Types of mutation
- 10 DNA repair and disorders
- 11 Human Chromosome and Chromosomal aberrations
- 12 DNA Recombination
- 13 References

Replication

In a cell, DNA replication begins at specific locations in the genome, called “**origins**”. Unwinding of DNA at the origin, and synthesis of new strands, forms a **replication fork**. In addition to DNA polymerase, the enzyme that synthesizes the new DNA by adding nucleotides matched to the template strand, a number of other proteins are associated with the fork and assist in the initiation and continuation of DNA synthesis. DNA replication can also be performed *in vitro* (outside a cell). DNA polymerases, isolated from cells, and artificial DNA primers are used to initiate DNA synthesis at known sequences in a template molecule. The polymerase chain reaction (**PCR**), a common laboratory technique, employs such artificial synthesis in a cyclic manner to amplify a specific target DNA fragment from a pool of DNA.^[1]



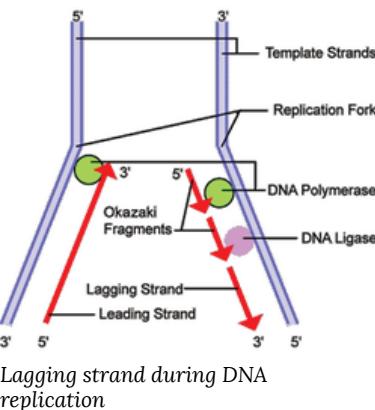
Schematic representation of the replication fork.
a: template, **b:** leading strand, **c:** lagging strand, **d:** replication fork, **e:** primer, **f:** Okazaki fragments

Leading strand

The leading strand template is the template strand of the DNA double helix that is oriented in a 3' to 5' manner. All DNA synthesis occurs 5'-3'. The original DNA strand must be read 3'-5' to produce a 5'-3' nascent strand. The leading strand is formed along the leading strand template as a polymerase “reads” the template DNA and continuously adds nucleotides to the 3' end of the elongating strand. This polymerase is DNA polymerase III (DNA Pol III) in prokaryotes and presumably Pol ε in eukaryotes.

Lagging strand

The lagging strand template is the coding strand of the DNA double helix that is oriented in a 5' to 3' manner. The newly made lagging strand still is synthesized 5'-3'. However, since the DNA is oriented in a manner that does not allow continual synthesis, only small sections can be read at a time. An RNA primer is placed on the DNA strand 3' to the origin of replication. Just as before, DNA Polymerase reads 3'-5' on the original DNA to produce a 5'-3' nascent strand. Polymerase reaches the origin of replication and stops replication until a new RNA primer is placed 3' to the last RNA primer. These fragments of DNA produced on the lagging strand are called Okazaki fragments. The orientation of the original DNA on the lagging strand prevents continual synthesis. As a result, replication of the lagging strand is more complicated than of the leading strand. On the lagging strand template, primase "reads" the DNA and adds RNA to it in short, separated segments. In eukaryotes, primase is intrinsic to **Pol α** . **DNA polymerase III** or **Pol δ** lengthens the primed segments, forming **Okazaki fragments**. Primer removal in eukaryotes is also performed by Pol δ . In prokaryotes, DNA polymerase I "reads" the fragments, removes the RNA using its flap endonuclease domain, and replaces the RNA nucleotides with DNA nucleotides (this is necessary because RNA and DNA use slightly different kinds of nucleotides). DNA ligase joins the fragments together.



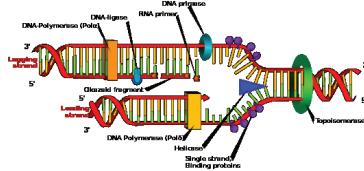
Lagging strand during DNA replication

Okazaki fragment

An **Okazaki fragment** is a relatively short fragment of DNA (with no RNA primer at the 5' terminus) created on the lagging strand during DNA replication. The lengths of Okazaki fragments are between 1,000 to 2,000 nucleotides long in *E. coli* and are generally between 100 to 200 nucleotides long in eukaryotes. It was originally discovered in 1968 by Reiji Okazaki, Tsuneko Okazaki, and their colleagues while studying replication of bacteriophage DNA in *Escherichia coli*.^{[2][3]}

Rate of replication

The rate of DNA replication in a living cell was first measured as the rate of phage T4 DNA elongation in phage-infected *E. coli*.^[4] During the period of exponential DNA increase at 37 °C, the rate was 749 nucleotides per second. The mutation rate per base pair per replication during phage T4 DNA synthesis is 1.7 per 10^8 .^[5] Thus semiconservative DNA replication is both rapid and accurate.



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.

Classification of DNA polymerase

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

1. Family A

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

2. Family B

In XPV patients, alternative error-prone polymerases, e.g., Pol ζ (zeta) (polymerase ζ is a B Family polymerase a complex of the catalytic subunit REV3L with Rev7, which associates with Rev1), are thought to be involved in mistakes that result in the cancer predisposition of these patients. The DNA polymerase which belongs to B family contain DTDS motif. The other members are Pol ϵ , Pol α , Pol δ .

3. Family C

Polymerases are the primary bacterial chromosomal replicative enzymes. DNA Polymerase III alpha subunit from E. coli is the catalytic subunit and possesses no known nuclease activity. A separate subunit, the epsilon subunit, possesses the 3'-5' exonuclease activity used for editing during chromosomal replication. Recent research has classified Family C polymerases as a subcategory of Family X.

4. Family D

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

5. Family X

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

6. Family Y

Y Polymerases differ from others in having a low fidelity on undamaged templates and in their ability to replicate through damaged DNA. Members of this family are hence called translesion synthesis (TLS) polymerases. Depending on the lesion, TLS polymerases can bypass the damage in an error-free or error-prone fashion, the latter resulting in elevated mutagenesis. Xeroderma pigmentosum variant (XPV) patients for instance have mutations in the gene encoding Pol η (eta), which is error-free for UV-lesions. Other members in humans are Pol ι (iota), Pol κ (kappa), and Rev1 (terminal deoxycytidyl transferase). In *E. coli*, two TLS polymerases, Pol IV (DINB) and Pol V (UmuD'2C), are known.

7. Family RT (reverse transcriptase)

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

Replication is semiconservative

The **Meselson and Stahl experiment** was an experiment by Matthew Meselson and Franklin Stahl in 1958 which supported the hypothesis that DNA replication was semiconservative.

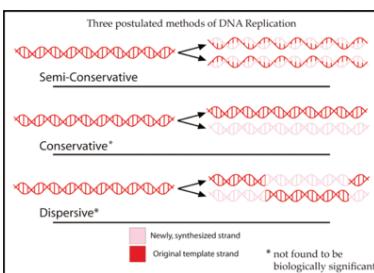
Semiconservative replication means that when the double stranded DNA helix was replicated, each of the two double stranded DNA helices consisted of one strand coming from the original helix and one newly synthesized. It has been called “the most beautiful experiment in biology.^[6]”

Three hypotheses had been previously proposed for the method of replication of DNA.

In the *semiconservative* hypothesis, proposed by Watson and Crick, the two strands of a DNA molecule separate during replication. Each strand then acts as a template for synthesis of a new strand.^[7]

The *conservative* hypothesis proposed that the entire DNA molecule acted as a template for synthesis of an entirely new one. According to this model, histone proteins bound to the DNA, distorting it in such a way as to expose both strands’ bases for hydrogen bonding.^[8]

The *dispersive* hypothesis is exemplified by a model proposed by Max Delbrück, which attempts to solve the problem of unwinding the two strands of the double helix by a mechanism that breaks the DNA backbone every 10 nucleotides or so, untwists the molecule, and attaches the old strand to the end of the newly synthesized one. This would synthesize the DNA in short pieces alternating from one strand to the other.^[9]



A summary of the three postulated methods of DNA synthesis

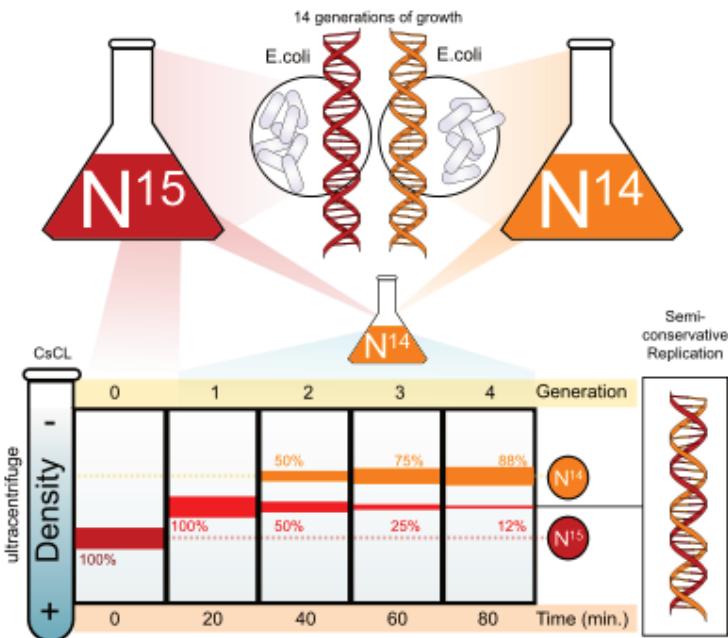
Each of these three models makes a different prediction about the distribution of the “old” DNA in molecules formed after replication. In the conservative hypothesis, after replication, one molecule is the entirely conserved “old” molecule, and the other is all newly synthesized DNA. The semiconservative hypothesis predicts that each molecule after replication will contain one old and one new strand. The dispersive model predicts that each strand of each new molecule will contain a mixture of old and new DNA.^[10]

The semi-conservative theory can be confirmed by making use of the fact that DNA is made up of nitrogen bases. Nitrogen has an isotope N15 (N14 is the most common isotope) called heavy nitrogen. The experiment that confirms the predictions of the semi-conservative theory^{[11][12]} makes use of this isotope and runs as follows: Bacterial (E coli) DNA is placed in a media containing heavy nitrogen(N15), which binds to the DNA, making it identifiable. Bacteria containing this DNA are then placed in a media with the presence of N14 and left to replicate only once. The new bases will contain nitrogen 14 while the originals will contain N15. The DNA is placed in test tubes containing caesium chloride (heavy compound) and centrifuged at 40,000 revolutions per minute. The caesium chloride molecules sink to the bottom of the test tubes creating a density gradient. The DNA molecules will position at their corresponding level of density (taking into account that N15 is more dense than N14). These test tubes are observed under ultraviolet rays. DNA appears as a fine layer in the test tubes at different heights according to their density. According to the semi-conservative theory, after one replication of DNA, we should obtain 2 hybrid (part N14 part N15) molecules from each original strand of DNA. This would appear as a single line in the test tube. This result would be the same for the dispersive theory. On the other hand, according to the conservative theory, we should obtain one original DNA duplex and a completely new one i.e. two fine lines in the test tube placed separately one from the other. Up to this point, either the semi-conservative or the dispersive theories could be truthful, as experimental evidence confirmed that only one line appeared.

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

Genetic evidence

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



Nitrogen is a major constituent of DNA. ¹⁴N is by far the most abundant isotope of nitrogen, but DNA with the heavier (but non-radioactive) ¹⁵N isotope is also functional.

E. coli were grown for several generations in a medium with ¹⁵N. When DNA is extracted from these cells and centrifuged on a salt density gradient, the DNA separates out at the point at which its density equals that of the salt solution. The DNA of the cells grown in ¹⁵N medium had a higher density than cells grown in normal ¹⁴N medium. After that, E. coli cells with only ¹⁵N in their DNA were transferred to a ¹⁴N medium and were allowed to divide; the progress of cell division was monitored by measuring the optical density of the cell suspension.

DNA was extracted periodically and was compared to pure ¹⁴N DNA and ¹⁵N DNA. After one replication, the DNA was found to have close to the intermediate density. Since conservative replication would result in equal amounts of DNA of the higher and lower densities (but no DNA of an intermediate density), conservative

replication was excluded. However, this result was consistent with both semiconservative and dispersive replication. Semiconservative replication would result in double-stranded DNA with one strand of ^{15}N DNA, and one of ^{14}N DNA, while dispersive replication would result in double-stranded DNA with both strands having mixtures of ^{15}N and ^{14}N DNA, either of which would have appeared as DNA of an intermediate density.

The authors continued to sample cells as replication continued. DNA from cells after two replications had been completed was found to consist of equal amounts of DNA with two different densities, one corresponding to the intermediate density of DNA of cells grown for only one division in ^{14}N medium, the other corresponding to DNA from cells grown exclusively in ^{14}N medium. This was inconsistent with dispersive replication, which would have resulted in a single density, lower than the intermediate density of the one-generation cells, but still higher than cells grown only in ^{14}N DNA medium, as the original ^{15}N DNA would have been split evenly among all DNA strands. The result was consistent with the semiconservative replication hypothesis [11]

Replication in prokaryote

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

Primase

In bacteria, primase binds to the DNA helicase forming a complex called the primosome. Primase is activated by DNA helicase where it then synthesizes a short RNA primer approximately 11 ± 1 nucleotides long, to which new nucleotides can be added by DNA polymerase.

Primosome

A primosome is a protein complex responsible for creating RNA primers on single stranded DNA during DNA replication. Primosomes are nucleoproteins assemblies that activate DNA replication forks. Their primary role is to recruit the replicative helicase onto single-stranded DNA. The “replication restart” primosome, defined in *Escherichia coli*, is involved in the reactivation of arrested replication forks.

Assembly of the *Escherichia coli* primosome requires six proteins, PriA, PriB, PriC, DnaB, DnaC, and DnaT, acting at a primosome assembly site (pas) on an SSBcoated single-stranded (8s) DNA. Assembly is initiated by interactions of PriA and PriB with ssDNA and the pas. PriC, DnaB, DnaC, and DnaT then act on the PriAPriB-DNA complex to yield the primosome.

The primosome consists of **seven proteins**: **DnaG** primase, **DnaB** helicase, **DnaC** helicase assistant, **DnaT**, **PriA**, **Pri B**, and **PriC**. The primosome is utilized once on the leading strand of DNA and repeatedly, initiating each Okazaki fragment, on the lagging DNA strand. Initially the complex formed by **PriA**, **PriB**, and **PriC** binds to DNA. Then the DnaB-DnaC helicase complex attaches along with **DnaT**. This structure is referred to as the **pre-primosome**. Finally, **DnaG** will bind to the pre-primosome forming a complete primosome. The primosome attaches 1-10 RNA nucleotides to the single stranded DNA creating a DNA-RNA hybrid. This sequence of RNA is used as a primer to initiate DNA polymerase III. The RNA bases are ultimately replaced with DNA bases by RNase H nuclease (eukaryotes) or DNA polymerase I nuclease (prokaryotes). DNA Ligase then acts to join the two ends together.

Elongation of DNA strand

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

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

DNA is read in the $3' \rightarrow 5'$ direction, therefore, nucleotides are synthesized (or attached to the template strand) in the $5' \rightarrow 3'$ direction. However, one of the parent strands of DNA is $3' \rightarrow 5'$ while the other is $5' \rightarrow 3'$. To solve this, replication occurs in opposite directions. Heading towards the replication fork, the **leading strand**

is synthesized in a continuous fashion, only requiring one primer. On the other hand, the **lagging strand**, heading away from the replication fork, is synthesized in a series of **short fragments known as Okazaki fragments**, consequently requiring many primers. The RNA primers of Okazaki fragments are subsequently degraded by RNase H and DNA Polymerase I (exonuclease), and the gap (or nicks) are filled with deoxyribonucleotides and sealed by the enzyme ligase.^[13]

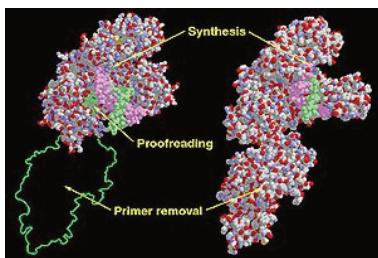
Termination

Termination of DNA replication in *E. coli* is completed through the use of termination sequences and the **Tus** protein. Tus is a sequence-specific DNA binding protein that promotes termination in prokaryotic DNA replication. In *E. Coli*, Tus binds to ten closely related 23 basepair binding sites encoded in the bacterial chromosome. These sites, called **Ter sites**, are designated **TerA, TerB, ..., TerJ**. The binding sites are asymmetric, such that when a Tus-Ter complex (Tus protein bound to a Ter site) is encountered by a replication fork from one direction, the complex is dissociated and replication continues (permissive). When encountered from the other direction, however, the Tus-Ter complex provides a much larger kinetic barrier and halts replication (non-permissive). The multiple Ter sites in the chromosome are oriented such that the two oppositely moving replication forks are both stalled in the desired termination region.

DNA polymerase in prokaryote

In Prokaryotic there are 5 kind of DNA polymerases:

Pol I: implicated in DNA repair; has 5'→3' polymerase activity, and both 3'→5' exonuclease (proofreading) and 5'→3' exonuclease activity (RNA primer removal).



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

Pol II: involved in repairing damaged DNA; has 3'→5' exonuclease activity. The enzyme is 90 kDa in size and is coded by the *polB* gene. DNA Pol II can synthesize DNA new base pairs at an average rate of between 40 and 50 nucleotides/second.

Pol III: the main polymerase in bacteria (responsible for elongation); has 3'→5' exonuclease activity (proofreading). The replisome is composed of the following:

- 2 DNA Pol III enzymes, made up of α , ϵ and θ subunits.
- the α subunit synthesizes the RNA/DNA primer.
- the ϵ subunit synthesizes the leading strand.
- the θ subunit stimulates the ϵ subunit's proofreading.
- 2 β units which act as sliding DNA clamps, they keep the polymerase bound to the DNA.
- 2 τ units which acts to dimerize two of the core enzymes (α , ϵ , and θ subunits).
- 1 γ unit which acts as a clamp loader for the lagging strand Okazaki fragments, helping the two β subunits to form a unit and bind to DNA.
- The γ unit is made up of 5 γ subunits which include 3 γ subunits, 1 δ subunit, and 1 δ' subunit. The δ is involved in copying of the lagging strand

Pol IV: a Y-family DNA polymerase.

Pol V: a Y-family DNA polymerase; participates in bypassing DNA damage.

DNA Polymerase I or Pol I

DNA Polymerase I (or Pol I) is an enzyme that participates in the process of DNA replication in prokaryotes. It contains 928 amino acids, and is an example of a processive enzyme – it can sequentially catalyze multiple polymerisations. It was Discovered by **Arthur Kornberg in 1956**, it was the first known DNA polymerase (and, indeed, the first known of any kind of polymerase). It was initially characterized in *E. coli*, although it is ubiquitous in prokaryotes. In *E. coli* and many other bacteria, the gene which encodes **Pol I is known as polA**. Pol I possesses three enzymatic activities: (1) a 5' → 3' (forward) DNA polymerase activity, requiring a 3' primer site and a template strand; (2) a 3' → 5' (reverse) exonuclease activity that mediates proofreading; and (3) a 5' → 3' (forward) exonuclease activity mediating nick translation during DNA repair.

Klenow fragment

The 5' → 3' exonuclease activity of DNA polymerase I from *E. coli* makes it unsuitable for many applications, the Klenow fragment, which lacks this activity, can be very useful in research. The Klenow fragment is extremely useful for research-based tasks such as: (1) Synthesis of double-stranded DNA from single-stranded templates; (2) Filling in (meaning removal of overhangs to create blunt ends) recessed 3' ends of DNA fragments; (3) Digesting away protruding 3' overhangs; (4) Preparation of radioactive DNA probes. The Klenow fragment was also the original enzyme used for greatly amplifying segments of DNA in the polymerase chain reaction (PCR) process,

before being replaced by thermostable enzymes such as Taq polymerase.

Helicases and Topoisomerase

Many cellular processes (DNA replication, transcription, translation, recombination, DNA repair, ribosome biogenesis) involve the separation of nucleic acid strands. Helicases are often utilized to separate strands of a DNA double helix or a self-annealed RNA molecule using the energy from ATP hydrolysis, a process characterized by the breaking of hydrogen bonds between annealed nucleotide bases. They move incrementally along one nucleic acid strand of the duplex with a directionality and processivity specific to each particular enzyme. There are many helicases (**14 confirmed in E. coli, 24 in human cells**) resulting from the great variety of processes in which strand separation must be catalyzed.

Helicases adopt different structures and oligomerization states. Whereas DnaB-like helicases unwind DNA as donut-shaped hexamers, other enzymes have been shown to be active as monomers or dimers. Studies have shown that helicases may act passively, waiting for uncatalyzed unwinding to take place and then translocating between displaced strands,[1] or can play an active role in catalyzing strand separation using the energy generated in ATP hydrolysis. In the latter case, the helicase acts comparably to an active motor, unwinding and translocating along its substrate as a direct result of its ATPase activity. Helicases may process much faster in vivo than in vitro due to the presence of accessory proteins that aid in the destabilization of the fork junction. Defects in the gene that codes helicase cause Werner syndrome, a disorder characterized by the appearance of premature aging.^[14]

Superfamilies

Helicases have been classified in 5 superfamilies (SF1-SF5). All of the proteins bind ATP, and, as a consequence, all of them carry the

classical Walker A (phosphate-binding loop or P-loop) and Walker B (Mg^{2+} -binding aspartic acid) motifs.

Superfamily I: UvrD (E. coli, DNA repair), Rep (E. coli, DNA replication), PcrA (Staphylococcus aureus, recombination), Dda (bacteriophage T4, replication initiation), RecD (E. coli, recombinational repair), TraI (F-plasmid, conjugative DNA transfer). This family includes RNA helicases thought to be involved in duplex unwinding during viral RNA replication. Members of this family are found in positive-strand single-stranded RNA viruses from superfamily 1. This helicase has multiple roles at different stages of viral RNA replication, as dissected by mutational analysis.

Superfamily II: RecQ (E. coli, DNA repair), eIF4A (Baker's Yeast, RNA translation), WRN (human, DNA repair), NS3[5] (Hepatitis C virus, replication), TRCF (Mfd) (E.coli, transcription-repair coupling).

Superfamily III: LTag (Simian Virus 40, replication), E1 (human papillomavirus, replication), Rep (Adeno-Associated Virus, replication, viral integration, virion packaging). Superfamily 3 consists of helicases encoded mainly by small DNA viruses and some large nucleocytoplasmic DNA viruses.[6][7] Small viruses are very dependent on the host-cell machinery to replicate. SF3 helicase in small viruses is associated with an origin-binding domain. By pairing a domain that recognises the ori with a helicase, the virus can bypass the host-cell-based regulation pathway and initiate its own replication. The protein binds to the viral ori leading to origin unwinding. Cellular replication proteins are then recruited to the ori, and the viral DNA is replicated.

DnaB-like family: dnaB (E. coli, replication), gp41 (bacteriophage T4, DNA replication), T7gp4 (bacteriophage T7, DNA replication).

Rho-like family: Rho (E. coli, transcription termination). Note that these superfamilies do not subsume all possible helicases. For example, XPB and ERCC2 are helicases not included in any of the above families.

RNA Helicases

RNA Helicases and DNA Helicases can be found together in all the

Helicase Super Families except for SF6.^[15] However, not all RNA Helicases exhibit helicase activity as defined by enzymatic function, i.e., proteins of the Swi/Snf family. Although these proteins carry the typical helicase motifs, hydrolyze ATP in a nucleic acid-dependent manner, and are built around a helicase core, in general, no unwinding activity is observed.^[16]

RNA Helicases that do exhibit unwinding activity have been characterized by at least two different mechanisms: canonical duplex unwinding and local strand separation. Canonical duplex unwinding is the stepwise directional separation of a duplex strand, as described above, for DNA unwinding. However, local strand separation occurs by a process wherein the helicase enzyme is loaded at any place along the duplex. This is usually aided by a single-stranded region of the RNA, and the loading of the enzyme is accompanied with ATP binding.^[17] Once the helicase and ATP are bound, local strand separation occurs, which requires binding of ATP but not the actual process of ATP hydrolysis.^[18] Presented with fewer base pairs the duplex then dissociates without further assistance from the enzyme. This mode of unwinding is used by DEAD-box helicases.^[19]

Topoisomerase

Topoisomerases are enzymes that unwind and wind DNA, in order for DNA to control the synthesis of proteins, and to facilitate DNA replication. The double-helical configuration that DNA strands naturally reside in makes them difficult to separate, and yet they must be separated by helicase proteins if other enzymes are to transcribe the sequences that encode proteins, or if chromosomes are to be replicated. In so-called circular DNA, in which double helical DNA is bent around and joined in a circle, the two strands are topologically linked, or knotted. Otherwise, identical loops of DNA having different numbers of twists are topoisomers, and cannot be interconverted by any process that does not involve the breaking of DNA strands. Topoisomerases catalyze and guide the unknotting or unkinking of DNA^[3] by creating transient breaks in the DNA using a conserved Tyrosine as the catalytic residue. The insertion of viral

DNA into chromosomes and other forms of recombination can also require the action of topoisomerases.

Topoisomerases can fix these topological problems and are separated into two types separated by the number of strands cut in one round of action.^[20] Both these classes of enzyme utilize a conserved tyrosine. However these enzymes are structurally and mechanistically different.

- Type I topoisomerase cuts one strand of a DNA double helix, relaxation occurs, and then the cut strand is reannealed. Cutting one strand allows the part of the molecule on one side of the cut to rotate around the uncut strand, thereby reducing stress from too much or too little twist in the helix. Such stress is introduced when the DNA strand is “supercoiled” or uncoiled to or from higher orders of coiling. Type I topoisomerases are subdivided into two subclasses: type IA topoisomerases, which share many structural and mechanistic features with the type II topoisomerases, and type IB topoisomerases, which utilize a controlled rotary mechanism. Examples of type IA topoisomerases include topo I and topo III. In the past, type IB topoisomerases were referred to as eukaryotic topo I, but IB topoisomerases are present in all three domains of life. It is interesting to note that type IA topoisomerases form a covalent intermediate with the 5' end of DNA, while the IB topoisomerases form a covalent intermediate with the 3' end of DNA. Recently, a type IC topoisomerase has been identified, called topo V. While it is structurally unique from type IA and IB topoisomerases, it shares a similar mechanism with type IB topoisomerase.
- Type II topoisomerase cuts both strands of one DNA double helix, passes another unbroken DNA helix through it, and then reanneals the cut strand. It is also split into two subclasses: type IIA and type IIB topoisomerases, which share similar structure and mechanisms. Examples of type IIA topoisomerases include eukaryotic topo II, E. coli gyrase, and

E. coli topo IV. Examples of type IIB topoisomerase include topo VI.

Topoisomerase	IA	IB	IIA	IIB
Metal Dependence	Yes	No	Yes	Yes
ATP Dependence	No	No	Yes	Yes
Single- or Double-Stranded cleavage?	SS	SS	DS	DS
Cleavage Polarity	5'	3'	5'	5'
Change in L	±1	±N	±2	±2

Both type I and type II topoisomerases change the linking number (L) of DNA. Type IA topoisomerases change the linking number by one, type IB and type IC topoisomerases change the linking number by any integer, while type IIA and type IIB topoisomerases change the linking number by two.

Many drugs operate through interference with the topoisomerases. The broad-spectrum fluoroquinolone antibiotics act by disrupting the function of bacterial type II topoisomerases. Some chemotherapy drugs work by interfering with topoisomerases in cancer cells:

type 1 is inhibited by irinotecan and topotecan.

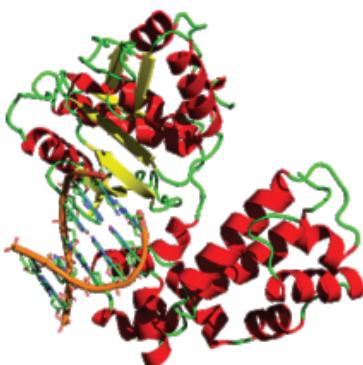
type 2 is inhibited by etoposide (VP-16), teniposide and HU-331, a quinolone synthesized from cannabidiol.

Topoisomerase I is the antigen recognized by Anti Scl-70 antibodies in scleroderma.

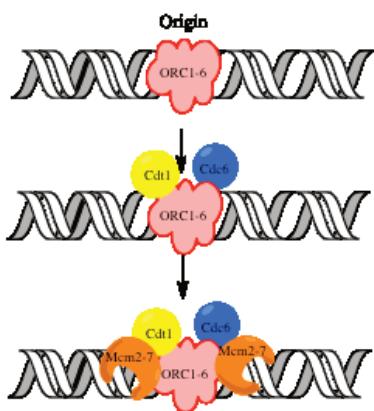
These small molecule inhibitors act as efficient anti-bacterial and anti-cancer agents by hijacking the natural ability of topoisomerase to create breaks in chromosomal DNA. These breaks in DNA accumulate, ultimately leading to programmed cell death, or apoptosis.

Replication in Eukaryote

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



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



Pre-RC assembly involves the assembly of the ORC subunits, Cdc6 and Cdt1 and the Mcm2-7 complex

DNA replication in eukaryotes occurs only in the S phase of the cell cycle. However, pre-initiation occurs in the G1 phase. Thus, the separation of pre-initiation and activation ensures that the origin can only fire once per cell cycle. Due to the sheer size of chromosomes in eukaryotes, eukaryotic chromosomes contain **multiple origins of replication**. Some origins are well characterized, such as the **autonomously replicating**

sequences (ARS) of yeast while other eukaryotic origins, particularly those in metazoa, can be found in spans of thousands of basepairs.^[21]

Eukaryotic DNA polymerase

There are at least 15 known Eukaryotic DNA polymerase:

POLA1, POLA2: Pol α (also called RNA primase): forms a complex with a small catalytic (PriS) and a large noncatalytic (PriL) subunit, with the Pri subunits acting as a primase (synthesizing an RNA primer), and then with DNA Pol α elongating that primer with DNA nucleotides. After around 20 nucleotides[3] elongation is taken over by Pol ϵ (on the leading strand) and δ (on the lagging strand).

POLB: Pol β : Implicated in repairing DNA, in base excision repair and gap-filling synthesis.

POLG, POLG2: Pol γ : Replicates and repairs mitochondrial DNA and has proofreading 3'->5' exonuclease activity.

POLD1, POLD2, POLD3, POLD4: Pol δ : Highly processive and has proofreading 3'->5' exonuclease activity. Thought to be the main polymerase involved in lagging strand synthesis, though there is still debate about its role.

POLE, POLE2, POLE3: Pol ϵ : Also highly processive and has proofreading 3'->5' exonuclease activity. Highly related to pol δ , and thought to be the main polymerase involved in leading strand synthesis[5], though there is again still debate about its role.

POLH, POLI, POLK,: η , ι , κ , and Rev1 are Y-family DNA polymerases and Pol ζ is a B-family DNA polymerase. These polymerases are involved in the bypass of DNA damage.

There are also other eukaryotic polymerases known, which are not as well characterized:

POLQ: ' θ

POLL: λ ϕ σ

POLM: μ

None of the eukaryotic polymerases can remove primers (5'->3' exonuclease activity); that function is carried out by other enzymes. Only the polymerases that deal with the elongation (γ , δ and ϵ) have proofreading ability (3'->5' exonuclease).

Preparation in G1 phase

The first step in DNA replication is the formation of the pre-initiation replication complex (the pre-RC). The formation of this complex occurs in two stages. The first stage requires that there is no CDK activity. This can only occur in early G1. The formation of the pre-RC is known as licensing, but a licensed pre-RC cannot initiate replication in the G1 phase. Current models hold that it begins with the binding of the origin recognition complex (ORC) to the origin. This complex is a hexamer of related proteins and remains bound to the origin, even after DNA replication occurs. Furthermore, ORC is the functional analogue of prokaryotic DnaA. Following the binding of ORC to the origin, Cdc6/Cdc18 and Cdt1 coordinate the loading of the MCM (Mini Chromosome Maintenance) complex to the origin by first binding to ORC and then binding to the MCM complex. The MCM complex is thought to be the major DNA helicase in eukaryotic organisms. Once binding of MCM occurs, a fully licensed pre-RC exists.

DNA Replication occurs during the S phase

Activation of the complex occurs in S-phase and requires Cdk2-Cyclin E and Ddk. The activation process begins with the addition of Mcm10 to the pre-RC, which displaces Cdt1. Following this, Ddk phosphorylates Mcm3-7, which activates the helicase. It is believed that ORC and Cdc6/18 are phosphorylated by Cdk2-Cyclin E. Ddk and the Cdk complex then recruits another protein called Cdc45, which then recruits all of the DNA replication proteins to the replication fork. At this stage the origin fires and DNA synthesis begins. Activation of a new round of replication is prevented through the actions of the cyclin dependent kinases and a protein known as geminin. Geminin binds to Cdt1 and sequesters it. It is a periodic protein that first appears in S-phase and is degraded in late M-phase, possibly through the action of the anaphase promoting complex (APC). In addition, phosphorylation of Cdc6/18 prevent

it from binding to the ORC (thus inhibiting loading of the MCM complex) while the phosphorylation of ORC remains unclear. Cells in the G0 stage of the cell cycle are prevented from initiating a round of replication because the Mcm proteins are not expressed.

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

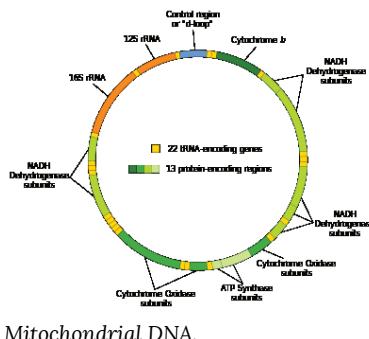
Pol α forms a complex with a small catalytic (PriS) and a large noncatalytic (PriL) subunit, with the Pri subunits acting as a primase (synthesizing an RNA primer), and then with DNA Pol α elongating that primer with DNA nucleotides. After around 20 nucleotides elongation is taken over by Pol ϵ (on the leading strand) and δ (on the lagging strand).

Pol δ : Highly processive and has proofreading 3'->5' exonuclease activity. Thought to be the main polymerase involved in leading strand synthesis, though there is still debate about its role.

Pol ϵ : Also highly processive and has proofreading 3'->5' exonuclease activity. Highly related to pol δ , and thought to be the main polymerase involved in lagging strand synthesis, though there is again still debate about its role.^[22]

Replication in mitochondria

Nuclear and mitochondrial DNA are thought to be of separate evolutionary origin, with the mtDNA being derived



Mitochondrial DNA.

from the circular genomes of the bacteria that were engulfed by the early ancestors of today's eukaryotic cells. This theory is called the endosymbiotic theory. Each mitochondrion is estimated to contain 2-10 mtDNA copies. In the cells of extant organisms, the vast majority of the proteins present in the mitochondria (numbering

approximately 1500 different types in mammals) are coded for by nuclear DNA, but the genes for some of them, if not most, are thought to have originally been of bacterial origin, having since been transferred to the eukaryotic nucleus during evolution.

mtDNA is replicated by the DNA polymerase gamma complex which is composed of a 140 kDa catalytic DNA polymerase encoded by the POLG gene and a 55 kDa accessory subunit encoded by the POLG2 gene. During embryogenesis, replication of mtDNA is strictly down-regulated from the fertilized oocyte through the preimplantation embryo. At the blastocyst stage, the onset of mtDNA replication is specific to the cells of the trophectoderm. In contrast, the cells of the inner cell mass restrict mtDNA replication until they receive the signals to differentiate to specific cell types. D-loop replication is a process by which chloroplasts and mitochondria replicate their genetic material. An important component of understanding D-loop replication is that chloroplasts and mitochondria have a single circular chromosome like bacteria instead of the linear chromosomes found in eukaryotes. In many organisms, one strand of DNA in the plastid comprises heavier nucleotides (relatively more purines: adenine and guanine). This strand is called the H (heavy) strand. The L (light) strand comprises lighter nucleotides (pyrimidines: thymine and cytosine). Replication begins with replication of the heavy strand starting at the D-loop (also known as the control region). An origin of replication opens, and the heavy strand is replicated in one direction. After heavy strand replication has continued for some time, a new light strand is also synthesized, through the opening of another origin of replication. When diagramed, the resulting structure looks like the letter D. The D-loop region is important for phylogeographic studies. Because the region does not code for any genes, it is free to vary with only a few selective limitations on size and heavy/light strand factors. The mutation rate is among the fastest of anywhere in either the nuclear or mitochondrial genomes in animals. Mutations in the D-loop can effectively track recent and rapid

evolutionary changes such as within species and among very closely related species.^[23]

C₀t values

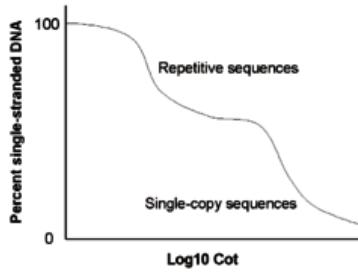
It was first developed and utilized by Roy Britten and his colleagues at the Carnegie Institution of Washington in the 1960s.^{[24][25]} Of particular note, it was through Cot analysis that the redundant (repetitive) nature of eukaryotic genomes was first discovered. Repeated sequences in DNA.^[26] However, it wasn't until the breakthrough DNA reassociation kinetics experiments of Britten and his colleagues that it was shown that not all DNA coded for genes. In fact, their experiments demonstrated that the majority of eukaryotic genomic DNA is composed of repetitive, non-coding elements. The amount of single and double-stranded DNA is measured by rapidly diluting the sample, which slows reassociation, and then binding the DNA to a hydroxylapatite column. The column is first washed with a low concentration of sodium phosphate buffer, which elutes the single-stranded DNA, and then with high concentrations of phosphate, which elutes the double stranded DNA. The amount of DNA in these two solutions is then measured using a spectrophotometer. Since a sequence of single-stranded DNA needs to find its complementary strand to reform a double helix, common sequences renature more rapidly than rare sequences. Indeed, the rate at which a sequence will reassociate is proportional to the number of copies of that sequence in the DNA sample. A sample with a highly-repetitive sequence will renature rapidly, while complex sequences will renature slowly. However, instead of simply measuring the percentage of double-stranded DNA versus time, the amount of renaturation is measured relative to a C0t value. The C0t value is the product of C0 (the initial concentration of DNA), t (time in seconds), and a constant that depends on the concentration of cations in the buffer. Repetitive

DNA will renature at low C_{Ot} values, while complex and unique DNA sequences will renature at high C_{Ot} values.

DNA repair

DNA damage, due to environmental factors and normal metabolic processes inside the cell, occurs at a rate of 1,000 to 1,000,000 molecular lesions per cell per day. While this constitutes only 0.000165% of the human genome's approximately 6 billion bases (**3 billion base pairs**), unrepaired lesions in critical genes (such as tumor suppressor genes) can impede a cell's ability to carry out its function and appreciably increase the likelihood of tumor formation.

The vast majority of DNA damage affects the primary structure of the double helix; that is, the bases themselves are chemically modified. These modifications can in turn disrupt the molecules' regular helical structure by introducing non-native chemical bonds or bulky adducts that do not fit in the standard double helix. Unlike proteins and RNA, DNA usually lacks tertiary structure and therefore damage or disturbance does not occur at that level. DNA is, however, supercoiled and wound around "packaging" proteins called histones (in eukaryotes), and both superstructures are vulnerable to the effects of DNA damage.^[27]



Repetitive DNA sequences renature at lower C_{Ot} values than single-copy sequences.

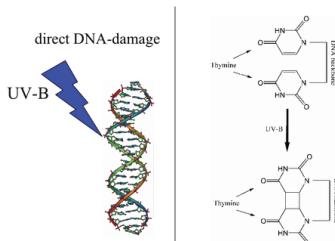
Types of DNA damage

There are five main types of damage to DNA due to endogenous cellular processes: (1) oxidation of bases [e.g. 8-oxo-7,8-dihydroguanine (8-oxoG)] and generation of DNA strand interruptions from reactive oxygen species; (2) alkylation of bases (usually methylation), such as formation of 7-methylguanine, 1-methyladenine, 6-O-Methylguanine; (3) hydrolysis of bases, such as deamination, depurination, and depyrimidination; (4) “bulky adduct formation” (i.e., benzo[a]pyrene diol epoxide-dG adduct); (5) mismatch of bases, due to errors in DNA replication, in which the wrong DNA base is stitched into place in a newly forming DNA strand, or a DNA base is skipped over or mistakenly inserted.

Damage caused by exogenous agents Damage caused by exogenous agents comes in many forms. Some examples are described below. UV-B light causes crosslinking between adjacent cytosine and thymine bases creating **pyrimidine dimers**. This is called direct DNA damage.

UV-A light creates mostly free radicals. The damage caused by free radicals is called indirect DNA damage.

Ionizing radiation such as that created by radioactive decay or in cosmic rays causes breaks in DNA strands. Low-level ionizing radiation may induce irreparable DNA damage (leading to replication and transcriptional errors needed for neoplasia or may trigger viral interactions) leading to pre-mature aging and cancer.



Direct DNA damage: The UV-photon is directly absorbed by the DNA (left). One of the possible reactions from the excited state is the formation of a thymine-thymine cyclobutane dimer (right). The direct DNA damage leads to sunburn, causing an increase in melanin production, thereby leading to a long-lasting tan. However, it is responsible for only 8% of all melanoma.

Thermal disruption at elevated temperature increases the rate of depurination (loss of purine bases from the DNA backbone) and single-strand breaks. For example, hydrolytic depurination is seen in the thermophilic bacteria, which grow in hot springs at 40–80 °C. The rate of depurination (300 purine residues per genome per generation) is too high in these species to be repaired by normal repair machinery, hence a possibility of an adaptive response cannot be ruled out.

Industrial chemicals also play very important role in DNA damage, such as vinyl chloride and hydrogen peroxide, and environmental chemicals such as polycyclic hydrocarbons found in smoke, soot and tar create a huge diversity of DNA adducts- ethenobases, oxidized bases, alkylated phosphotriesters and Crosslinking of DNA just to name a few. **UV damage, alkylation/methylation, X-ray damage and oxidative damage are examples of induced damage.** Spontaneous damage can include the loss of a base, deamination, sugar ring puckering and tautomeric shift.

Sources of damage

DNA damage can be subdivided into **two** main types:

Endogenous damage such as attack by reactive oxygen species produced from normal metabolic byproducts, especially the process of oxidative deamination, and this also includes base mismatches due to replication errors

Exogenous damage caused by external agents such as ultraviolet [UV 200–300 nm] radiation from the sun other radiation frequencies, including x-rays and gamma rays hydrolysis or thermal disruption certain plant toxins human-made mutagenic chemicals, especially aromatic compounds that act as DNA intercalating agents cancer chemotherapy and radiotherapy

Types of mutation

When DNA damages are repaired this can sometimes give rise to a simple one base-pair mutation, described here. (Deletions and translocations can also arise during repair)

Transition In molecular biology, a transition is a point mutation that changes a purine nucleotide to another purine ($A \leftrightarrow G$) or a pyrimidine nucleotide to another pyrimidine ($C \leftrightarrow T$). Approximately two out of three single nucleotide polymorphisms (SNPs) are transitions. Transitions can be caused by oxidative deamination and tautomerization. Although there are twice as many possible transversions, transitions appear more often in genomes, possibly due to the molecular mechanisms that generate them. 5-Methylcytosine is more prone to transition than unmethylated cytosine, due to spontaneous deamination. This mechanism is important because it dictates the rarity of CpG islands.

Transversion In molecular biology, transversion refers to the substitution of a purine for a pyrimidine or vice versa. It can only be reverted by a spontaneous reversion. Because this type of mutation changes the chemical structure dramatically, the consequences of this change tend to be more severe and less common than that of transitions. Transversions can be caused by ionizing radiation and alkylating agents.

DNA repair and disorders

Defects in the NER mechanism are responsible for squally several genetic disorders, including:

Xeroderma pigmentosum: hypersensitivity to sunlight/UV, resulting in increased skin cancer incidence and premature aging

Cockayne syndrome: hypersensitivity to UV and chemical agents

Trichothiodystrophy: sensitive skin, brittle hair and nails Mental retardation often accompanies the latter two disorders, suggesting increased vulnerability of developmental neurons.

Other DNA repair disorders include:

Werner's syndrome: premature aging and retarded growth

Bloom's syndrome: sunlight hypersensitivity, high incidence of malignancies (especially leukemias).

Ataxia telangiectasia: sensitivity to ionizing radiation and some chemical agents

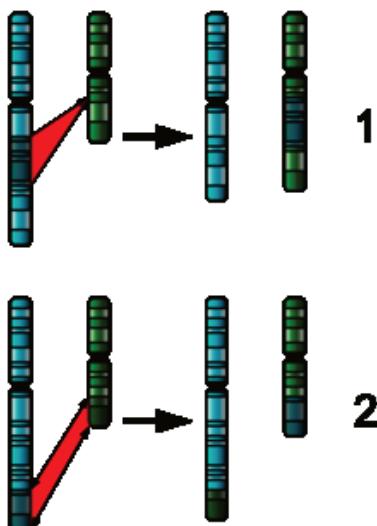
All of the above diseases are often called “**segmental progerias**” (“accelerated aging diseases”) because their victims appear elderly and suffer from aging-related diseases at an abnormally young age, while not manifesting all the symptoms of old age.

Other diseases associated with reduced DNA repair function include **Fanconi's anemia**, hereditary breast cancer and hereditary colon cancer.

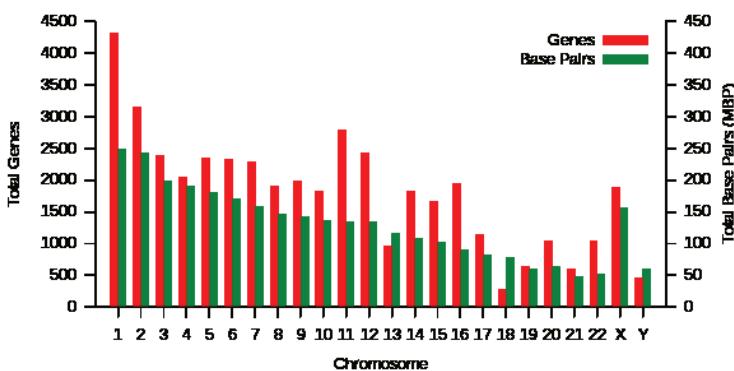
Human Chromosome and Chromosomal aberrations

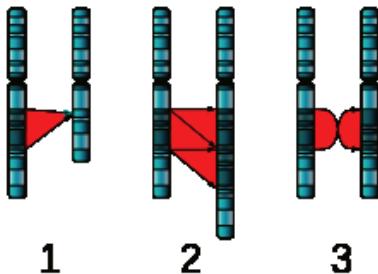
Chromosomes can be divided into two types—autosomes, and sex chromosomes. Certain genetic traits are linked to your sex, and are passed on through the sex chromosomes. The autosomes contain the rest of the genetic hereditary information. All act in the same way during cell division. Human cells have 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.



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

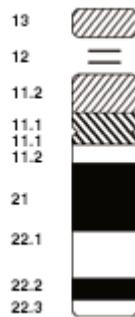




The three major single chromosome mutations; deletion (1), duplication (2) and inversion (3).

Chromosome	Genes	Total bases	Sequenced bases^[28]
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

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:



In Down syndrome, there are three copies of chromosome 21

- 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.^[29]
- 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.^[30]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.^[31]

DNA Recombination

Recombination is a process by which a molecule of nucleic acid (usually DNA, but can also be RNA) is broken and then joined to a different one (or in which genetic information is exchanged between two such molecules). Recombination ordinarily occurs between similar molecules of DNA, as in homologous recombination.

Recombination is a common method of DNA repair in both bacteria and eukaryotes. In eukaryotes, recombination also occurs in

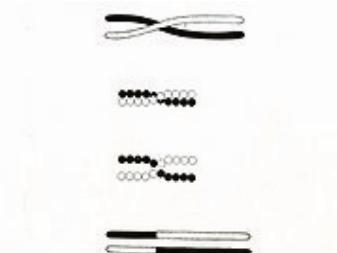


Fig. 61. Scheme to illustrate a method of crossing over of the chromosomes.

Thomas Hunt Morgan's illustration of crossing over (1916)

meiosis, where it facilitates informational exchange and/or chromosomal crossover. The crossover process leads to offspring's having different combinations of genes from those of their parents, and can occasionally produce new chimeric alleles. In organisms with an adaptive immune system, a type of genetic recombination called V(D)J recombination helps immune cells rapidly diversify to recognize and adapt to new pathogens. The shuffling of genes brought about by genetic recombination can have long term advantages, as it is a major engine of genetic variation and also allows sexually reproducing organisms to avoid Muller's ratchet, in which the genomes of an asexual population accumulate deleterious mutations in an irreversible manner. In genetic engineering, recombination can also refer to artificial and deliberate recombination of disparate pieces of DNA, often from different organisms, creating what is called recombinant DNA. A prime example of such a use of genetic recombination is gene targeting, which can be used to add, delete or otherwise change an organism's genes. This technique is important to biomedical researchers as it allows them to study the effects of specific genes. Techniques based on genetic recombination are also applied in protein engineering to develop new proteins of biological interest.^[32]

Chromosomal crossover in eukaryotes is an exchange of genetic material between homologous chromosomes. It can occur in one of the final phases of genetic recombination, which occurs during prophase I of meiosis (pachytene). The pairing of homologous chromosomes during meiosis (synapsis) begins before the synaptonemal complex develops, and is not completed until near the end of prophase I. Crossover usually occurs when matching regions on matching chromosomes break and then reconnect to the other chromosome.

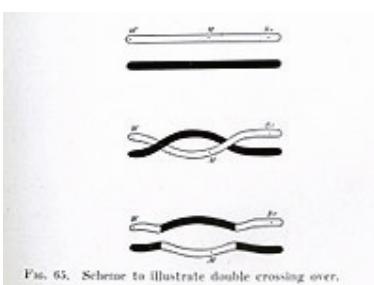
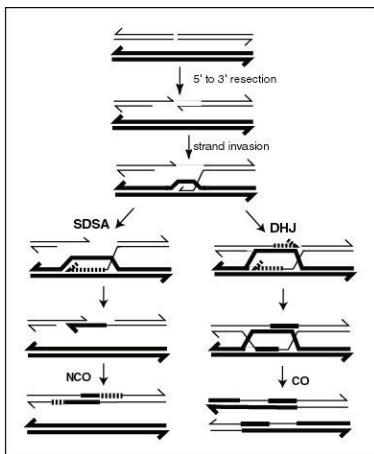


Fig. 65. Scheme to illustrate double crossing over.

Crossing over was described, in theory, by Thomas Hunt Morgan. He relied on the discovery of the Belgian Professor Frans Alfons Janssens of the University of Leuven who described the phenomenon in 1909 and had called it 'chiasmatypie'. The term chiasma is linked if not identical to chromosomal crossover. Morgan immediately saw the great importance of Janssens' cytological interpretation of chiasmata to the experimental results of his research on the heredity of *Drosophila*. The physical basis of crossing over was first demonstrated by Harriet Creighton and Barbara McClintock in 1931.

Meiotic recombination can be initiated by double-stranded breaks that can be introduced into the DNA by the Spo11 protein. In addition, meiotic recombination can be induced in response to spontaneous double strand breaks, possibly caused by reactive oxygen species, carried over from the prior round of synthesis.^[33] One or more exonucleases then digest the 5' ends generated by the double-stranded breaks to produce 3' single-stranded DNA tails (see lowest Figure in this section). The meiosis-specific recombinase Dmc1 and the general recombinase Rad51 coat the single-stranded DNA to form nucleoprotein filaments. The recombinases catalyze invasion of the opposite chromatid by the single-stranded DNA from one end of the break. Next, the 3' end of the invading DNA primes DNA synthesis, causing displacement of the complementary strand.



A current model of meiotic recombination, initiated by a double-strand break or gap, followed by pairing with an homologous chromosome and strand invasion to initiate the recombinational repair process. Repair of the gap can lead to crossover (CO) or non-crossover (NCO) of the flanking regions. CO recombination is thought to occur by the Double Holliday Junction (DHJ) model, illustrated on the right, above. NCO recombinants are thought to occur primarily by the Synthesis Dependent Strand Annealing (SDSA) model, illustrated on the left, above. Most recombination events appear to be the SDSA type.

Crossover recombinants are generated by a process in which the displaced complementary strand subsequently anneals to the single-stranded DNA generated from the other end of the initial double-stranded break (see DHJ pathway on the Figure). The structure that results is a cross-strand exchange, also known as a Holliday junction. The contact between two chromatids that will soon undergo crossing-over is known as a chiasma. The Holliday

junction is a tetrahedral structure which can be ‘pulled’ by other recombinases, moving it along the four-stranded structure (see Double Holliday Junction or DHJ in the Figure).

Gene conversion can result from the repair of a double strand break. Gene conversion involves the unidirectional transfer of genetic sequence information from a ‘donor’ sequence to a highly homologous ‘acceptor’ chromosome. Gene conversion usually occurs by Synthesis Dependent Strand Annealing (SDSA)^{[34][35][36]} illustrated in the lowest Figure in this section. In this model of SDSA DNA repair, a free strand of DNA from the end of a double strand break invades an homologous chromosome, extending itself by replication along the sequence on the complementary strand of DNA of the ‘donor’ chromosome. The extended strand is then retracted from the donor chromosome and pairs with the complementary sequence on the recipient chromosome in a region at the other end of the double strand break (needing about 25 to 50 base pairs of homology).^[34] This allows completion of healing of the double strand break by replication, to complete the duplex structure on the recipient chromosome, from information on the extended strand copied from the donor chromosome. The usual length of a gene conversion tract in mammals is between 200 to 1,000 base pairs.^[37]

During meiosis, gene conversion is most often associated with non-crossover of outside regions (e.g. the SDSA pathway shown in the Figure). Less frequently, gene conversion during meiosis is associated with crossover of outside regions and these events are usually generated by the DHJ pathway. Gene conversion without crossover occurs more frequently than crossover recombination during meiosis in many organisms, often by about a 2 to 1 ratio.^[38] During mitosis, gene conversion is almost the exclusive mode of double strand break repair by homologous recombination.^[36]

Studies of gene conversion have contributed to our understanding of the adaptive function of meiotic recombination. Since gene conversion in most species studied is more frequently of the non-crossover type,^[38] explanations for the adaptive function

of meiotic recombination that focus exclusively on the adaptive benefit of producing new genetic variation seem inadequate to explain the majority of recombination events during meiosis. However, the majority of meiotic recombination events can be explained by the proposal that they are an adaptation for repair of damages in the DNA that is to be passed on to gametes.^{[39][40]}

Genetic recombination is catalyzed by enzymes called recombinases. RecA, the chief recombinase found in Escherichia coli, is responsible for the repair of DNA double strand breaks (DSBs). In yeast and other eukaryotic organisms there are two recombinases required for repairing DSBs. The RAD51 protein is employed in both mitotic and meiotic recombination, whereas the DMC1 protein is specific to meiotic recombination.

Nonhomologous recombination Recombinational repair can infrequently occur between DNA sequences that contain no or little sequence homology. This is referred to as nonhomologous recombination.

References

1. ↑ DNA replication
2. ↑ Okazaki R, Okazaki T, Sakabe K, Sugimoto K. Mechanism of DNA replication possible discontinuity of DNA chain growth. An American scientist, by the last name Shandell, discovered this mechanism prior to Okazaki, however he was never accredited with the discovery since the head of his research team decided the discovery was an erroneous interpretation of test results. Jpn J Med Sci Biol. 1967 Jun;20(3):255-60.
3. ↑ Ogawa T, Okazaki T, Discontinuous DNA Replication. Annu. Rev. Biochem. 49:421-457, 1980
4. ↑ “DNA elongation rates and growing point distributions of wild-type phage T4 and a DNA-delay amber mutant”. J Mol Biol **106** (4): 963–81. 1976. doi:10.1016/0022-2836(76)90346-6.

- PMID789903.
5. ↑ Drake JW (1970) *The Molecular Basis of Mutation*. Holden-Day, San Francisco ISBN 0816224501 ISBN 978-0816224500
 6. ↑ John Cairns to Horace F Judson, in *The Eighth Day of Creation: Makers of the Revolution in Biology* (1979). Touchstone Books, ISBN 0-671-22540-5. 2nd edition: Cold Spring Harbor Laboratory Press, 1996 paperback: ISBN 0-87969-478-5.
 7. ↑ WATSON JD, CRICK FH (1953). "The structure of DNA". *Cold Spring Harbor Laboratory* **18**: 123–31. PMID13168976.
 8. ↑ Bloch DP (December 1955). "A POSSIBLE MECHANISM FOR THE REPLICATION OF THE HELICAL STRUCTURE OF DESOXYRIBONUCLEIC ACID". *Proc. Natl. Acad. Sci. U.S.A.* **41** (12): 1058–64. PMID16589796.
 9. ↑ Delbrück M (September 1954). "ON THE REPLICATION OF DESOXYRIBONUCLEIC ACID (DNA)". *Proc. Natl. Acad. Sci. U.S.A.* **40** (9): 783–8. PMID16589559.
 10. ↑ Delbrück, Max; Stent, Gunther S. (1957). "On the mechanism of DNA replication". in McElroy, William D.; Glass, Bentley. *A Symposium on the Chemical Basis of Heredity*. Johns Hopkins Pr.. pp. 699–736.
 11. ↑ Jump up to: **ab** Meselson, M. and Stahl, F.W. (1958). "The Replication of DNA in Escherichia coli". *PNAS* **44**: 671–82. doi:10.1073/pnas.44.7.671. PMID16590258.
[http://www.pubmedcentral.nih.gov/
articlerender.fcgi?tool=pubmed&pubmedid=16590258](http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=16590258).
 12. ↑ Meselson M, Stahl FW. Demonstration of the semiconservative mode of DNA duplication. In "Phage and the Origins of Molecular Biology" (editors Cairns J, Stent GS, Watson JD) pages 246–251 of Cold Spring Harbor Laboratory of Quantitative Biology, First edition (1966). ASIN: B00C2G89LM
 13. ↑ Prokaryotic DNA replication
 14. ↑ Helicase
 15. ↑ "RNA Helicases" Edited by Eckhard Jankowsky, RSC Publishing 2010

16. ↑ Trends Biochem Sci. 2010 Aug 31. RNA helicases at work: binding and rearranging. Jankowsky E. Center for RNA Molecular Biology & Department of Biochemistry, School of Medicine, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106, USA
17. ↑ Yang et al., DEAD-box proteins unwind duplexes by local strand separation, Mol. Cell 28 (2007), pp. 253–263
18. ↑ Liu et al., ATP hydrolysis is required for DEAD-box protein recycling but not for duplex unwinding, Proc. Natl. Acad. Sci. U. S. A. 105 (2008), pp. 20209–20214
19. ↑ Jarmoskaite, I. and Russell, R., (2010) DEAD-box proteins as RNA helicases and chaperones. WIREs: RNA, in press.
20. ↑ Wang JC (April 1991). “DNA topoisomerases: why so many?”. J. Biol. Chem. 266 (11): 6659–62. PMID1849888.
<http://www.jbc.org/cgi/pmidlookup?view=long&pmid=1849888>.
21. ↑ Eukaryotic DNA replication
22. ↑ http://en.wikipedia.org/w/index.php?title=Eukaryotic_DNA_replication&oldid=423827289
23. ↑ Mitochondrial DNA
24. ↑ Davidson EH, Britten RJ (1973) Organization, transcription, and regulation in the animal genome. Quart. Rev. Biol. 48: 565–613.
25. ↑ Britten RJ, Graham DE, Neufeld BR (1974). “Analysis of repeating DNA sequences by reassociation”. Meth. Enzymol. 29 (0): 363–418. doi:10.1016/0076-6879(74)29033-5. PMID4850571.
26. ↑ Science 161: 529–540.
27. ↑ DNA repair
28. ↑ Sequenced percentages are based on fraction of euchromatin portion, as the Human Genome Project goals called for determination of only the euchromatic portion of the genome. Telomeres, centromeres, and other heterochromatic regions have been left undetermined, as have a small number of

- unclonable gaps. See <http://www.ncbi.nlm.nih.gov/genome/seq/> for more information on the Human Genome Project.
- 29. ↑ Miller, Kenneth R. (2000). "9-3". *Biology* (5th ed.). Upper Saddle River, New Jersey: Prentice Hall. pp. 194–5. ISBN0-13-436265-9.
 - 30. ↑ European Chromosome 11 Network
 - 31. ↑ ORNL.gov, Exploring Genes & Genetic Disorders
 - 32. ↑ Genetic recombination
 - 33. ↑ Carofiglio F, Inagaki A, de Vries S, Wassenaar E, Schoenmakers S, Vermeulen C, van Cappellen WA, Sleddens-Linkels E, Grootegoed JA, Te Riele HP, de Massy B, Baarends WM. (2013). SPO11-independent DNA repair foci and their role in meiotic silencing. *PLoS Genet.* 9(6):e1003538. doi: 10.1371/journal.pgen.1003538. PMID 23754961
 - 34. ↑ Jump up to: **ab** Allers T, Lichten M. (2001). Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* 106(1):47–57. PMID 11461701
 - 35. ↑ McMahill MS, Sham CW, Bishop DK. (2007). Synthesis-dependent strand annealing in meiosis. *PLoS Biol.* 5(11):e299. PMID 17988174 PMCID: PMC2062477.
 - 36. ↑ Jump up to: **ab** Andersen SL, Sekelsky J. (2010). Meiotic versus mitotic recombination: two different routes for double-strand break repair: the different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes. *Bioessays.* 32(12):1058–66. doi: 10.1002/bies.201000087. Review. PMID 20967781
 - 37. ↑ Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 8(10):762–75. Review. PMID 17846636
 - 38. ↑ Jump up to: **ab** Whitehouse, HLK. *Genetic Recombination.(see Table 38)* New York: Wiley; 1982. ISBN 0471102059 ISBN 978-0471102052
 - 39. ↑ Hörandl E. (2013). *Meiosis and the Paradox of Sex in Nature, Meiosis*, Dr. Carol Bernstein (Ed.), ISBN 978-953-51-1197-9,

InTech, DOI: 10.5772/56542. Available from:
<http://www.intechopen.com/books/meiosis/meiosis-and-the-paradox-of-sex-in-nature>

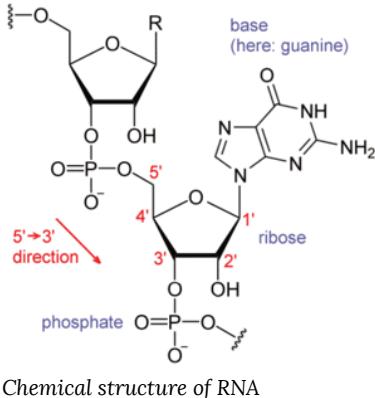
40. ↑ Bernstein H, BernsteinC and Michod RE. (2011). Meiosis as an Evolutionary Adaptation for DNA Repair. Chapter 19 in DNA Repair. Inna Kruman editor. InTech Open Publisher. DOI: 10.5772/25117 <http://www.intechopen.com/books/dna-repair/meiosis-as-an-evolutionary-adaptation-for-dna-repair>

I4.

Ribonucleic acid is popularly known as RNA. RNA is one of the three major macromolecules (along with DNA and proteins) that are essential for all known forms of life. 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). 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.^[1]

Contents

- 1 History of RNA
- 2 Types of RNA
 - 2.1 RNAs involved in protein synthesis
 - 2.1.1 Messenger RNA
 - 2.1.2 Ribosomal RNA
 - 2.1.3 Transfer RNA
 - 2.2 RNAs involved in post-transcriptional
 - 2.2.1 Small nuclear ribonucleic acid (snRNA)
 - 2.2.2 Small nucleolar RNAs (snoRNAs)
 - 2.2.3 Ribonuclease P (RNase P)
 - 2.2.4 Telomerase RNA
 - 2.3 Small interfering RNA (siRNA)
- 3 Structure of RNA
- 4 RNA base
- 5 DNA vs RNA
- 6 snRNA and snRNPs
- 7 RNA as an enzyme
- 8 Ribonuclease
- 9 RNAi
 - 9.1 RNAi and Disease control
- 10 MicroRNA (miRNA)
- 11 Facts to be remembered
- 12 References



History of RNA

Nucleic acids were discovered in 1868 by Friedrich Miescher, who called the material ‘nuclein’ since it was found in the nucleus.^[2] 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.^[3] 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.^[4] 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,^[5] 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.^{[6][7]}

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.^[8]

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.^{[9][10]}

At about the same time, 22 nt long RNAs, now called microRNAs, were found to have a role in the development of *C. elegans*.^[11] 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.^[12]

Types of RNA

RNAs involved in protein synthesis

Messenger RNA

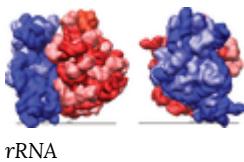
Messenger RNA (mRNA) is a molecule of RNA encoding a chemical "blueprint" for a protein product. mRNA is transcribed from a DNA template, and carries coding information to the sites of protein synthesis: the ribosomes. Here, the nucleic acid polymer is translated into a polymer of amino acids: a protein. In mRNA as in DNA, genetic information is encoded in the sequence of nucleotides arranged into codons consisting of three bases each. Each codon encodes for a specific amino acid, except the stop codons that terminate protein synthesis. This process requires two other types of RNA: transfer RNA (tRNA) mediates recognition of the codon and provides the corresponding amino acid, while ribosomal RNA (rRNA) is the central component of the ribosome's protein manufacturing machinery.



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

Ribosomal RNA

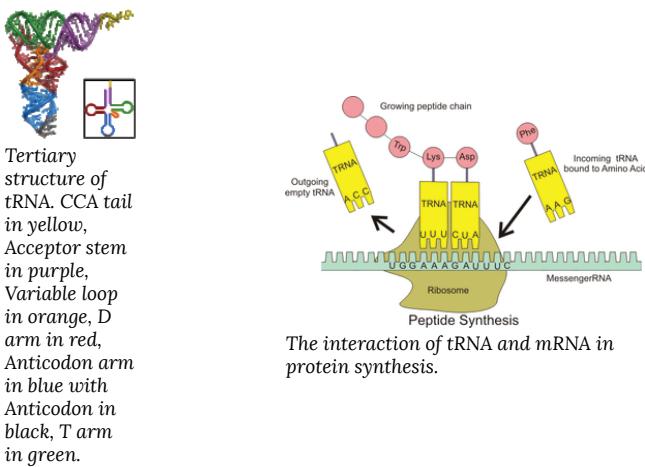
Ribosomal ribonucleic acid (rRNA) is the RNA component of the ribosome, the organelle that is the site of protein synthesis in all living cells. Ribosomal RNA provides a mechanism for decoding mRNA into amino acids and interacts with tRNAs during translation by providing peptidyl transferase activity. The tRNAs bring the necessary amino acids corresponding to the appropriate mRNA codon.



Transfer RNA

Transfer RNA (tRNA) is a small RNA molecule (usually about 73–95 nucleotides) that transfers a specific active amino acid to a growing

polypeptide chain at the ribosomal site of protein synthesis during translation. It has a 3' terminal site for amino acid attachment. This covalent linkage is catalyzed by an aminoacyl tRNA synthetase. It also contains a three base region called the anticodon that can base pair to the corresponding three base codon region on mRNA. Each type of tRNA molecule can be attached to only one type of amino acid, but because the genetic code contains multiple codons that specify the same amino acid, tRNA molecules bearing different anticodons may also carry the same amino acid.



RNAs involved in post-transcriptional

Small nuclear ribonucleic acid (snRNA)

Small nuclear ribonucleic acid (snRNA) is a class of small RNA molecules that are found within the nucleus of eukaryotic cells. They are transcribed by RNA polymerase II or RNA polymerase III and are involved in a variety of important processes such as

RNA splicing (removal of introns from hnRNA), regulation of transcription factors (7SK RNA) or RNA polymerase II (B2 RNA), and maintaining the telomeres. They are always associated with specific proteins, and the complexes are referred to as small nuclear ribonucleoproteins (snRNP) or sometimes as snurps. These elements are rich in uridine content.

Small nucleolar RNAs (snoRNAs)

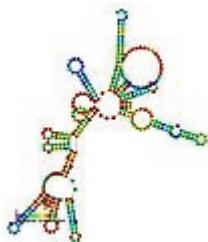
Small nucleolar RNAs (snoRNAs) are a class of small RNA molecules that primarily guide chemical modifications of other RNAs, mainly ribosomal RNAs, transfer RNAs and small nuclear RNAs. There are two main classes of snoRNA, the C/D box snoRNAs which are associated with methylation, and the H/ACA box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

After transcription, nascent rRNA molecules (termed pre-rRNA) are required to undergo a series of processing steps in order to generate the mature rRNA molecule. Prior to cleavage by exo- and endonucleases the pre-rRNA undergoes a complex pattern of nucleoside modifications. These include methylations and pseudouridylation, guided by snoRNAs. Methylation is the attachment or substitution of a methyl group onto various substrates. The rRNA of humans contain approximately 115 methyl group modifications. The majority of these are 2' O-ribose-methylations (where the methyl group is attached to the ribose group). Pseudouridylation is the conversion (isomerisation) of the nucleoside uridine to a different isomeric form pseudouridine(Ψ). Mature human rRNAs contain approximately 95 Ψ modifications. Each snoRNA molecule acts as a guide for only one (or two) individual modifications in a target RNA. In order to carry out modification, each snoRNA associates with at least four protein molecules in an RNA/protein complex referred to as a small

nucleolar ribonucleoprotein (snoRNP). The proteins associated with each RNA depend on the type of snoRNA molecule (see snoRNA guide families below). The snoRNA molecule contains an antisense element (a stretch of 10-20 nucleotides) which are base complementary to the sequence surrounding the base (nucleotide) targeted for modification in the pre-RNA molecule. This enables the snoRNP to recognise and bind to the target RNA. Once the snoRNP has bound to the target site the associated proteins are in the correct physical location to catalyse the chemical modification of the target base.^[13]

Ribonuclease P (RNase P)

Ribonuclease P (RNase P) is a type of Ribonuclease which cleaves RNA. RNase P is unique from other RNases in that it is a ribozyme – a ribonucleic acid that acts as a catalyst in the same way that a protein based enzyme would. Its function is to cleave off an extra, or precursor, sequence of RNA on tRNA molecules



Predicted secondary
structure and sequence
conservation of
RNaseP_bact_a

Telomerase RNA

Telomerase RNA component, also known as TERC, is an RNA gene found in eukaryotes, that is a component of telomerase used to extend telomeres. Telomerase RNAs differ greatly in sequence and structure between vertebrates, ciliates and yeasts, but they share a 5' pseudoknot structure close to the template sequence. The vertebrate telomerase RNAs have a 3' H/ACA snoRNA-like domain.

Telomerase RNA component		==Regulatory RNAs==	==Antisense RNA==	==Micro (miRNAs)
3D representation of part of the telomerase RNA component. This is the solution structure of the P2b-P3 pseudoknot from human telomerase RNA				

Small interfering RNA (siRNA)

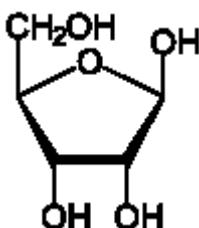
Small interfering RNA (siRNA), sometimes known as short interfering RNA or silencing RNA, is a class of double-stranded RNA molecules, 20-25 nucleotides in length, that play a variety of roles in biology. Most notably, siRNA is involved in the RNA interference (RNAi) pathway, where it interferes with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

of these pathways is only now being elucidated. siRNAs were first discovered by David Baulcombe's group at the Sainsbury Laboratory in Norwich, England, as part of post-transcriptional gene silencing (PTGS) in plants.

IDENTIFICATION OF SMALL-INTERFERING RNA

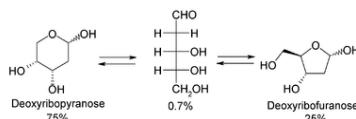
By using mapping studies, it showed that target transcript cleavage corresponded to regions that they are complementary to dsRNA and occurred at 21–22-nt intervals which looks like the same size of dsRNA-derived small RNAs. To test if small RNAs mediate RISC activity, 21– 22-nt RNA duplexes with symmetric 2-nt 3' overhangs were synthesized to mimic dsRNA- processing products. Certainly, these synthetic oligonucleotides induced target mRNA cleavage at sites corresponding to the middle of small RNA. Therefore, these dsRNA-derived small RNAs were designated as small-interfering RNAs (siRNAs). Nevertheless, a most important contribution was made by demonstration of target transcript silencing through transfection of synthetic siRNAs into mammalian cells. This work provided the foundation for numerous RNAi-based applications, including powerful “shut down” function strategy and a new potential therapeutics. As such, these milestone studies provided an spectacular example of the importance of basic science, including traditional biochemistry, in generating new areas for biomedical applications.

Structure of RNA



Ribose.

Each nucleotide in RNA contains a ribose sugar, with



Chemical equilibrium of deoxyribose in solution.

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.

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.

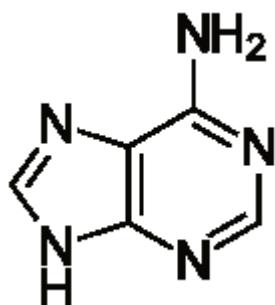
Ribose is an aldopentose, that is a monosaccharide containing five carbon atoms that, in its open chain form, has an aldehyde functional group at one end. In the conventional numbering scheme for monosaccharides, the carbon atoms are numbered from C1' (in

the aldehyde group) to C5'. The deoxyribose derivative, found in DNA, differs from ribose by having a hydrogen atom in place of the hydroxyl group in carbon C2'. Like many monosaccharides, ribose occurs in water as the linear form H-(C=O)-(CHOH)₄-H and any of two ring forms: ribofuranose ("C3'-endo"), with a five-membered ring, and ribopyranose ("C2'-endo"), with a six-membered ring. The ribofuranose form is predominant in aqueous solution. The "D-" in the name D-ribose refers to the stereochemistry of the chiral carbon atom farthest away from the aldehyde group (C4'). In D-ribose, as in all D-sugars, this carbon atom has the same configuration as in D-glyceraldehyde. Ribose comprises the backbone of RNA, a biopolymer that is the basis of genetic transcription. It is related to deoxyribose, as found in DNA. Once phosphorylated, ribose can become a subunit of ATP, NADH, and several other compounds that are critical to metabolism.^[14]

RNA base

adenine (A)

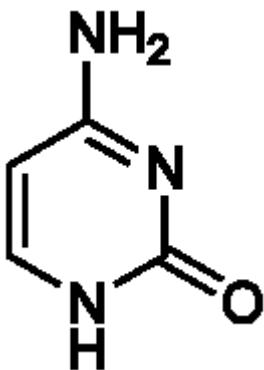
Adenine (A,) is a nucleobase (a purine derivative) with a variety of roles in biochemistry including cellular respiration, in the form of both the energy-rich adenosine triphosphate (ATP) and the cofactors nicotinamide adenine dinucleotide (NAD) and flavin adenine dinucleotide (FAD), and protein synthesis, as a chemical component of DNA and RNA. The shape of adenine is complementary to either thymine in DNA or uracil in RNA.



Adenine

cytosine (C)

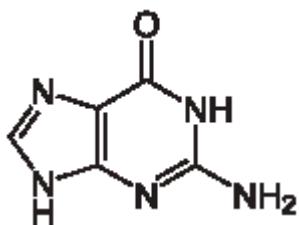
Cytosine (C) is one of the four main bases found in DNA and RNA, along with adenine, guanine, and thymine (uracil in RNA). It is a pyrimidine derivative, with a heterocyclic aromatic ring and two substituents attached (an amine group at position 4 and a keto group at position 2). The nucleoside of cytosine is cytidine.



Cytosine

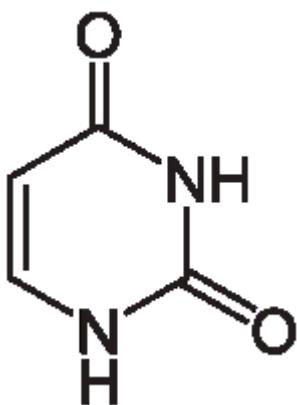
guanine (G)

Guanine (G) is one of the four main nucleobases found in the nucleic acids DNA and RNA, the others being adenine, cytosine, and thymine (uracil in RNA). In DNA, guanine is paired with cytosine. With the formula C5H5N5O, guanine is a derivative of purine, consisting of a fused pyrimidine-imidazole ring system with conjugated double bonds.



Guanine

uracil (U)



Uracil

Found in RNA, it base-pairs with adenine and replaces thymine during DNA transcription. Methylation of uracil produces thymine. It turns into thymine to protect the DNA and to improve the efficiency of DNA replication.

Uracil can base-pair with any of the bases, depending on how the molecule arranges itself on the helix, but readily pairs with adenine because the methyl group is repelled into a fixed position.^[15] Uracil pairs with adenine through hydrogen bonding. Uracil is the hydrogen bond acceptor and can form two hydrogen bonds. Uracil can also bind with a ribose sugar to form the ribonucleoside uridine. When a phosphate attaches to uridine, uridine 5'-monophosphate is produced.^[16]

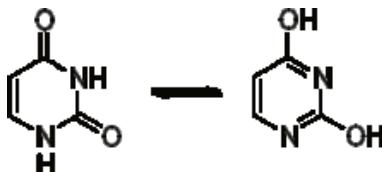
Uracil undergoes amide-imidic acid tautomeric shifts because any nuclear instability the molecule may have from the lack of formal aromaticity is compensated by the cyclic-amidic stability. The amide tautomer is referred to as the lactam structure, while the imidic acid tautomer is referred to as the lactim structure. These tautomeric forms are predominant at pH 7. The lactam structure is the most common form of uracil.

Uracil also recycles itself to form nucleotides by undergoing a series of phosphoribosyltransferase reactions. Degradation of uracil produces the substrates aspartate, carbon dioxide, and ammonia.



Oxidative degradation of uracil produces urea and maleic acid in the presence of H_2O_2 and Fe^{2+} or in the presence of diatomic oxygen and Fe^{2+} .

Uracil is a weak acid; the first site of ionization of uracil is not known.^[17] The negative charge is placed on the oxygen anion and



Uracil tautomers: Amide or lactam structure (left) and imide or lactim structure (right)

produces a pK_a of less than or equal to 12. The basic $pK_a = -3.4$, while the acidic $pK_a = 9.38$. In the gas phase, uracil has 4 sites that are more acidic than water.^[18]

Uracil is a common and naturally occurring pyrimidine derivative. Originally discovered in 1900, it was isolated by hydrolysis of yeast nuclein that was found in bovine thymus and spleen, herring sperm, and wheat germ. It is a planar, unsaturated compound that has the ability to absorb light.

DNA vs RNA

RNA and DNA are both nucleic acids, but differ in three main ways.

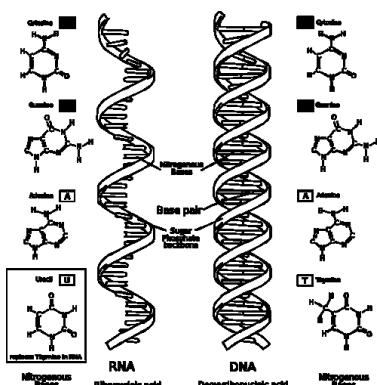
First, unlike DNA, which is, in general, double-stranded, RNA is a single-stranded molecule in many of its biological roles and has a much shorter chain of nucleotides.

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.

Fourth, DNA is generally stable in alkaline conditions while RNA is unstable during alkaline condition.

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



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

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.

snRNA and snRNPs

snRNPs (pronounced “snurps”), or small nuclear ribonucleoproteins, are RNA-protein complexes that combine with unmodified pre-mRNA and various other proteins to form a spliceosome, a large RNA-protein molecular complex upon which splicing of pre-mRNA occurs. The action of snRNPs is essential to the removal of introns from pre-mRNA, a critical aspect of post-transcriptional modification of RNA, occurring only in the nucleus of eukaryotic cells. The two essential components of snRNPs are protein molecules and RNA. The RNA found within each snRNP particle is known as small nuclear RNA, or snRNA, and is usually about 150 nucleotides in length. The snRNA component of the snRNP gives specificity to individual introns by “recognizing” the sequences of critical splicing signals at the 5' and 3' ends and branch site of introns. The snRNA in snRNPs is similar to ribosomal RNA in that it directly incorporates both an enzymatic and a structural role. SnRNPs were discovered by Michael R. Lerner and Joan A. Steitz. Thomas R. Cech and Sidney Altman also played a role in the discovery, winning the Nobel Prize for Chemistry in 1989 for their independent discoveries that RNA can act as a catalyst in cell development.

At least five different kinds of snRNPs join the spliceosome to participate in splicing. They can be visualized by gel electrophoresis

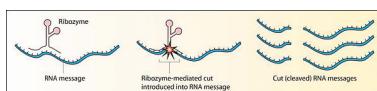
and are known individually as: U1, U2, U4, U5, and U6. Their snRNA components are known, respectively, as: U1 snRNA, U2 snRNA, U4 snRNA, U5 snRNA, and U6 snRNA. In the mid-1990s, it was discovered that a variant class of snRNPs exists to help in the splicing of a class of introns found only in metazoans, with highly-conserved 5' splice sites and branch sites. This variant class of snRNPs includes: U11 snRNA, U12 snRNA, U4atac snRNA, and U6atac snRNA. While different, they perform the same functions as do U1, U2, U4, and U6, respectively.^[19]

The completed core snRNP-snurportin 1 complex is transported into the nucleus via the protein importin β . Inside the nucleus, the core snRNPs appear in the Cajal bodies, where final assembly of the snRNPs take place. This consists of additional proteins and other modifications specific to the particular snRNP (U1, U2, U4, U5). The biogenesis of the U6 snRNP occurs in the nucleus although large amounts of free U6 are found in the cytoplasm. The LSm ring may assemble first, and then associate with the U6 snRNA.

Small nuclear ribonucleic acid (snRNA) is a class of small RNA molecules that are found within the nucleus of eukaryotic cells. They are transcribed by RNA polymerase II or RNA polymerase III and are involved in a variety of important processes such as RNA splicing (removal of introns from hnRNA), regulation of transcription factors (7SK RNA) or RNA polymerase II (B2 RNA), and maintaining the telomeres. They are always associated with specific proteins, and the complexes are referred to as small nuclear ribonucleoproteins (snRNP) or sometimes as snurps. These elements are rich in uridine content. A large group of snRNAs are known as small nucleolar RNAs (snoRNAs). These are small RNA molecules that play an essential role in RNA biogenesis and guide chemical modifications of ribosomal RNAs (rRNAs) and other RNA genes (tRNA and snRNAs). They are located in the nucleolus and the Cajal bodies of eukaryotic cells (the major sites of RNA synthesis).

RNA as an enzyme

Before the discovery of ribozymes, enzymes, which are defined as catalytic proteins,^[20] were the only known biological catalysts. In



Schematic showing ribozyme cleavage of RNA.

1967, Carl Woese, Francis Crick, and Leslie Orgel were the first to suggest that RNA could act as a catalyst. This idea was based upon the discovery that RNA can form complex secondary structures.^[21] The first ribozymes were discovered in the 1980s by Thomas R. Cech, who was studying RNA splicing in the ciliated protozoan *Tetrahymena thermophila* and Sidney Altman, who was working on the bacterial RNase P complex. These ribozymes were found in the intron of an RNA transcript, which removed itself from the transcript, as well as in the RNA component of the RNase P complex, which is involved in the maturation of pre-tRNAs. In 1989, Thomas R. Cech and Sidney Altman won the Nobel Prize in Chemistry for their “discovery of catalytic properties of RNA.”^[22] The term ribozyme was first introduced by Kelly Kruger *et al.* in 1982 in a paper published in the journal *Cell*.^[23]

RNA enzymes, or ribozymes, are still found in today's DNA-based life and could be examples of living fossils. Ribozymes play vital roles, such as those in the ribosome, which is vital for protein synthesis. Many other ribozyme functions exist, for example: the Hammerhead ribozyme performs self-cleavage^[24] and an RNA polymerase ribozyme can autocatalyse its own synthesis.^[25]

Among the enzymatic properties important for the beginning of life are:

- The ability to self-duplicate, or duplicate other RNA molecules. Relatively short RNA molecules that can duplicate others have been artificially produced in the lab. The shortest was 165-base long, though it has been estimated that only part of the bases

were crucial for this function. One version, 189-base long, had fidelity of 98.9%,^[26] which would mean it would make an exact copy of an RNA molecule as long as itself in one of every eight copies. This 189 base pair ribozyme could polymerize a template of at most 14 nucleotides in length, which is too short for replication, but a promising lead for further investigation. The longest primer extension by a ribozyme polymerase was 20 bases.^[27]

- The ability to catalyze simple chemical reactions which would enhance the creation of molecules which are building blocks of RNA molecules—i.e., a strand of RNA which would make creating more strands of RNA easier. Relatively short RNA molecules with such abilities have been artificially formed in the lab.^{[28][29]}
- The ability to catalyse the formation of peptide bonds, in order to produce short peptides, or—eventually—full proteins. This is done in modern cells by ribosomes, a complex of two large RNA molecules known as rRNA and many proteins. The two rRNA molecules are thought to be responsible for its enzymatic activity. A much shorter RNA molecule has been formed in lab with the ability to form peptide bonds, and it has been suggested that rRNA has evolved from a similar molecule.^[30] It has also been suggested that amino acids may have initially been complexed with RNA molecules as cofactors enhancing or diversifying their enzymatic capabilities, before evolving to the more complex peptides. mRNA may have evolved from such RNA molecules, and tRNA from RNA molecules which had catalyzed amino acid transfer to them.^[31]

Although most ribozymes are quite rare in the cell, their roles are sometimes essential to life. For example, the functional part of the ribosome, the molecular machine that translates RNA into proteins, is fundamentally a ribozyme, composed of RNA tertiary structural motifs that are often coordinated to metal ions such as Mg²⁺ as cofactors. There is no requirement for divalent cations in a five-

nucleotide RNA that can catalyze trans-phenylalanination of a four-nucleotide substrate which has three base complementary sequence with the catalyst. The catalyst and substrate were devised by truncation of the C3 ribozyme.

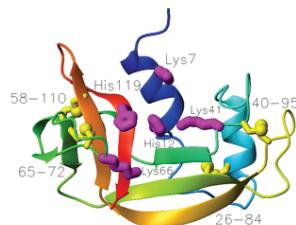
Ribonuclease

Ribonuclease (commonly abbreviated RNase) is a type of nuclease that catalyzes the degradation of RNA into smaller components. Ribonucleases can be divided into endoribonucleases and exoribonucleases, and comprise several sub-classes within the EC 2.7 (for the phosphorolytic enzymes) and 3.1 (for the hydrolytic enzymes) classes of enzymes.

Endoribonuclease is a ribonuclease, endonuclease. It

cleaves either single or double stranded RNA depending on the enzyme. Example includes both single proteins like RNase III, RNase A, RNase T1 and RNase H but also, complexes of proteins like RNase P and the RNA-induced silencing complex.

Exoribonuclease An exoribonuclease is an exonuclease ribonuclease, which are enzymes that degrade RNA by removing terminal nucleotides from either the 5' end or 3' end of the RNA molecule. Enzymes that remove nucleotides from the 5' end are called 5'-3' exoribonucleases and enzymes that remove nucleotides from the 3' end are called 3'-5' exoribonucleases.

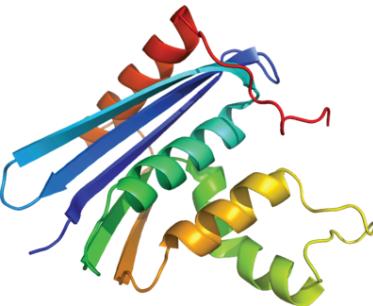


Labeled ribbon diagram della ribonuclease A pancreatica bovina (PDB accession code 7RSA). The backbone ribbon is colored from blue (N-terminus) to red (C-terminus). The side chains of the four disulfide-bonded cysteines are shown in yellow, with their sulfur atoms highlighted as small spheres. Residues important for catalysis are shown in magenta.

Exoribonucleases can use either water to cleave the nucleotide-nucleotide bond (which is called hydrolytic activity) or inorganic phosphate (which is called phosphorolytic activity). Hydrolytic exoribonucleases are classified under EC number 3.1 and phosphorolytic exoribonucleases under EC number 2.7.7. As the phosphorolytic enzymes use inorganic phosphate to cleave bonds they release nucleotide disphosphates), whereas the hydrolytic enzymes (which use water) release nucleotide monophosphates). Exoribonucleases exist in all kingdoms of life, the bacteria, archaea and eukaryotes. Exoribonucleases are involved in the degradation of many different RNA species, including messenger RNA, transfer RNA, ribosomal RNA and miRNA. Exoribonucleases can be single proteins (like RNase D or RNase PH) but also can be complexes of multiple proteins, like the exosome complex (in which four of the major exoribonuclease families are represented)^[32]

RNase A

RNase A is a relatively small protein (124 residues, ~13.7 kDa). It can be characterized as a two-layer $\alpha + \beta$ protein that is folded in half to resemble a taco, with a deep cleft for binding the RNA substrate. The first layer is composed of three alpha helices (residues 3-13, 24-34 and 50-60) from the N-terminal half of the protein. The second layer consist of three β -hairpins (residues 61-74, 79-104 and 105-124 from the C-terminal half) arranged in two β -sheets. The hairpins 61-74 and 105-124 form a four-stranded, antiparallel β -sheet that lies on helix 3 (residues 50-60). The longest β -hairpin 79-104 mates with a short β -strand (residues 42-45) to form a three-stranded, antiparallel β -sheet that lies on helix 2 (residues 24-34). RNase A has four disulfide bonds in its native state: Cys26-Cys84, Cys58-110, Cys40-95 and Cys65-72. The first two (26-84 and 58-110) are essential for conformational folding; each joins an alpha helix of



Structure the E.coli RNase H

the first layer to a beta sheet of the second layer, forming a small hydrophobic core in its vicinity. The latter two disulfide bonds (40-95 and 65-72) are less essential for folding; either one can be reduced (but not both) without affecting the native structure under physiological conditions. These disulfide bonds connect loop segments and are relatively exposed to solvent. Interestingly, the 65-72 disulfide bond has an extraordinarily high propensity to form, significantly more than would be expected from its loop entropy, both as a peptide and in the full-length protein. This suggests that the 61-74 β -hairpin has a high propensity to fold conformationally. RNase A is a basic protein ($pI = 8.63$); its many positive charges are consistent with its binding to RNA (a poly-anion). More generally, RNase A is unusually polar or, rather, unusually lacking in hydrophobic groups, especially aliphatic ones. This may account for its need of four disulfide bonds to stabilize its structure. The low hydrophobic content may also serve to reduce the physical repulsion between highly charged groups (its own and those of its substrate RNA) and regions of low dielectric constant (the nonpolar residues). The N-terminal α -helix of RNase A (residues 3-13) is connected to the rest of RNase A by a flexible linker (residues 16-23). As shown by F. M. Richards, this linker may be cleaved by subtilisin between residues 20 and 21 without causing the N-terminal helix to dissociate from the rest of RNase A. The peptide-protein complex is called RNase S, the peptide (residues 1-20) is called the S-peptide and the remainder (residues 21-124) is called the S-protein. The dissociation constant of the S-peptide for the S-protein is roughly 30 pM; this tight binding can be exploited for protein purification by attaching the S-peptide to the protein of interest and passing a mixture over an affinity column with bound S-protein. [A smaller C-peptide (residues 1-13) also works.] The RNase S model system has also been used for studying protein folding by coupling folding and association. The S-peptide was the first peptide from a native protein shown to have (flickering) secondary structure in isolation (by Klee and Brown in 1967).^[33]

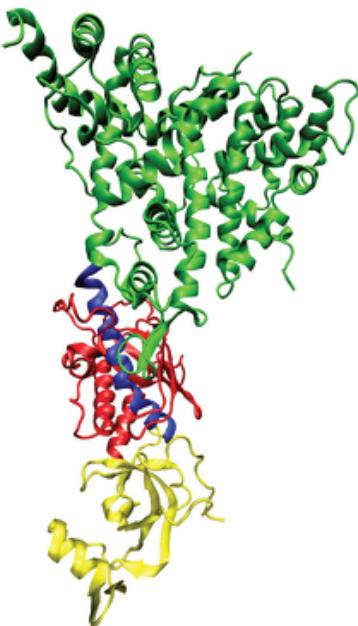
RNase H

In a molecular biology laboratory, as RNase H specifically degrades the RNA in RNA:DNA hybrids and will not degrade DNA or unhybridized RNA, it is commonly used to destroy the RNA template after first-strand complementary DNA (cDNA) synthesis by reverse transcription, as well as procedures such as nuclease protection assays. RNase H can also be used to degrade specific RNA strands when the cDNA oligo is hybridized, such as the removal of the poly(A) tail from mRNA hybridized to oligo(dT), or the destruction of a chosen non-coding RNA inside or outside the living cell. To terminate the reaction, a chelator, such as EDTA, is often added to sequester the required metal ions in the reaction mixture. The enzyme RNase H is a non-specific endonuclease and catalyzes the cleavage of RNA via a hydrolytic mechanism. Members of the RNase H family can be found in nearly all organisms, from archaea and prokaryota and eukaryota.

The 3-D structure of RNase H commonly consists of a 5-stranded β -sheet surrounded by a distribution of α -helices. In some RNase H, such as the one found in HIV-1, the enzyme is missing one of the helices known as the C-helix, a positively charged α -helix whose protrusive shape increases substrate binding capacity. The active site of the enzyme is centered around a conserved DEDD motif (composed of residues: D443, E478, D498, and D549) which performs the hydrolysis of the RNA substrate. A magnesium ion is commonly used as a cofactor during the hydrolysis step. It is also a potential but unconfirmed mechanism in which multiple ions are necessary for to perform the hydrolysis. The enzyme also contains a nucleic acid binding cleft about 60 Å in length that can encompass a region of 18 bound RNA/DNA base pairs. RNase H's ribonuclease activity cleaves the 3'-O-P bond of RNA in a DNA/RNA duplex to produce 3'-hydroxyl and 5'-phosphate terminated products. In DNA replication, RNase H is responsible for removing the RNA primer, allowing completion of the newly synthesized DNA.^[34]

RNAi

The discovery of RNAi was preceded first by observations of transcriptional inhibition by antisense RNA expressed in transgenic plants,^[36] and more directly by reports of unexpected outcomes in experiments performed by plant scientists in the United States and the Netherlands in the early 1990s.^[37] In an attempt to alter flower colors in petunias, researchers introduced additional copies of a gene encoding chalcone synthase, a key enzyme for flower pigmentation into petunia plants of normally pink or violet flower color. The overexpressed gene was expected to result in darker flowers, but instead produced less pigmented, fully or partially white flowers, indicating that the activity of chalcone synthase had been substantially decreased; in fact, both the endogenous genes and the transgenes were downregulated in the white flowers. Soon after, a related event termed *quelling* was noted in the fungus *Neurospora crassa*,^[38] although it was not immediately recognized as related. Further investigation of the phenomenon in plants indicated that the down regulation was due to post-transcriptional inhibition of gene expression via an increased rate of mRNA degradation.^[39] This



The dicer protein from *Giardia intestinalis*, which catalyzes the cleavage of dsRNA to siRNAs. The RNase domains are colored green, the PAZ domain yellow, the platform domain red, and the connector helix blue.^[35]

phenomenon was called co-suppression of gene expression, but the molecular mechanism remained unknown.^[40] Not long after, plant virologists working on improving plant resistance to viral diseases observed a similar unexpected phenomenon. While it was known that plants expressing virus-specific proteins showed enhanced tolerance or resistance to viral infection, it was not expected that plants carrying only short, non-coding regions of viral RNA sequences would show similar levels of protection. Researchers believed that viral RNA produced by transgenes could also inhibit viral replication.^[42] The reverse experiment, in which short sequences of plant genes were introduced into viruses, showed that the targeted gene was suppressed in an infected plant. This phenomenon was labeled “virus-induced gene silencing” (VIGS), and the set of such phenomena were collectively called post transcriptional gene silencing.^[43] After these initial observations in plants, many laboratories around the world searched for the occurrence of this phenomenon in other organisms.^{[44][45]} Craig C. Mello and Andrew Fire’s 1998 *Nature* paper reported a potent gene silencing effect after injecting double stranded RNA into *Caenorhabditis elegans*. In investigating the regulation of muscle protein production, they observed that neither mRNA nor antisense RNA injections had an effect on protein production, but double-stranded RNA successfully silenced the targeted gene. As a result of this work, they coined the term RNAi. Fire and Mello’s discovery was particularly notable because it represented the first identification of the causative agent for the phenomenon. Fire and Mello were awarded the Nobel Prize in Physiology or Medicine in 2006 for their work.

RNAi is an RNA-dependent gene silencing process that is controlled by the RNA-induced silencing complex (RISC) and is initiated by short double-stranded RNA molecules in a cell's cytoplasm, where they interact with the catalytic RISC component argonaute. When the dsRNA is exogenous



Example petunia plants in which genes for pigmentation are silenced by RNAi. The left plant is wild-type; the right plants contain transgenes that induce suppression of both transgene and endogenous gene expression, giving rise to the unpigmented white areas of the flower.^[4]

(coming from infection by a virus with an RNA genome or laboratory manipulations), the RNA is imported directly into the cytoplasm and cleaved to short fragments by the enzyme Dicer. The initiating dsRNA can also be endogenous (originating in the cell), as in pre-microRNAs expressed from RNA-coding genes in the genome. The primary transcripts from such genes are first processed to form the characteristic stem-loop structure of pre-miRNA in the nucleus, then exported to the cytoplasm to be cleaved by Dicer. Thus, the two dsRNA pathways, exogenous and endogenous, converge at the RISC complex.^[46]

dsRNA cleavage Endogenous dsRNA initiates RNAi by activating the ribonuclease protein Dicer, which binds and cleaves double-stranded RNAs (dsRNAs) to produce double-stranded fragments of 20–25 base pairs with a 2 nucleotide overhang at 3' end. Bioinformatics studies on the genomes of multiple organisms suggest this length maximizes target-gene specificity and minimizes non-specific effects.^[10] These short double-stranded fragments are called small interfering RNAs (siRNAs). These siRNAs are then separated into single strands and integrated into an active RISC complex. After integration into the RISC, siRNAs base-pair to their target mRNA and induce cleavage of the mRNA, thereby preventing it from being used as a translation template.

RISC RNA-Induced Silencing Complex, or RISC, is a multiprotein complex that incorporates one strand of a small interfering RNA (siRNA) or micro RNA (miRNA). RISC uses the siRNA or miRNA as a template for recognizing complementary mRNA.

When it finds a complementary strand, it activates RNase and cleaves the RNA. This process is important both in gene regulation by microRNAs and in defense against viral infections, which often use double-stranded RNA as an infectious vector. The active components of an RNA-induced silencing complex (RISC) are endonucleases called argonaute proteins, which cleave the target mRNA strand complementary to their bound siRNA. As the fragments produced by dicer are double-stranded, they could each in theory produce a functional siRNA. However, only one of the two strands, which is known as the guide strand, binds the argonaute protein and directs gene silencing. The other anti-guide strand or passenger strand is degraded during RISC activation. **Dicer** Dicer is an endoribonuclease in the RNase III family that cleaves double-stranded RNA (dsRNA) and pre-microRNA (miRNA) into short double-stranded RNA fragments called small interfering RNA (siRNA) about 20-25 nucleotides long, usually with a two-base overhang on the 3' end. Dicer contains two RNase III domains and one PAZ domain; the distance between these two regions of the molecule is determined by the length and angle of the connector helix and determines the length of the siRNAs it produces. Dicer catalyzes the first step in the RNA interference pathway and initiates formation of the RNA-induced silencing complex (RISC), whose catalytic component argonaute is an endonuclease capable of degrading messenger RNA (mRNA) whose sequence is complementary to that of the siRNA guide strand.^[47] The human version of this gene is DICER1.

RNAi and Disease control

It may be possible to exploit RNA interference in therapy. Although it is difficult to introduce long dsRNA strands into mammalian cells due to the interferon response, the use of short interfering RNA mimics has been more successful. Among the first applications to

reach clinical trials were in the treatment of macular degeneration and respiratory syncytial virus, RNAi has also been shown to be effective in the reversal of induced liver failure in mouse models.

Other proposed clinical uses center on antiviral therapies, including topical microbicide treatments that use RNAi to treat infection (at Harvard University Medical School; in mice, so far) by herpes simplex virus type 2 and the inhibition of viral gene expression in cancerous cells, knockdown of host receptors and coreceptors for HIV, the silencing of hepatitis A and hepatitis B genes, silencing of influenza gene expression, and inhibition of measles viral replication. Potential treatments for neurodegenerative diseases have also been proposed, with particular attention being paid to the polyglutamine diseases such as Huntington's disease. RNA interference is also often seen as a promising way to treat cancer by silencing genes differentially upregulated in tumor cells or genes involved in cell division. A key area of research in the use of RNAi for clinical applications is the development of a safe delivery method, which to date has involved mainly viral vector systems similar to those suggested for gene therapy.

MicroRNA (miRNA)

MicroRNAs were discovered in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum during a study of the gene lin-14 in *C. elegans* development. They found that LIN-14 protein abundance was regulated by a short RNA product encoded by the lin-4 gene. A 61 nucleotide precursor from lin-4 gene matured to a 22 nucleotide RNA containing sequences partially complementary to multiple sequences in the 3' UTR of the lin-14 mRNA. This complementarity was sufficient and necessary to inhibit the translation of lin-14 mRNA into LIN-14 protein. Retrospectively, the lin-4 small RNA was the first microRNA to be identified, though at the time, it was

thought to be a nematode idiosyncrasy. Only in 2000 was a second RNA characterized: let-7, which repressed lin-41, lin-14, lin-28, lin-42, and daf-12 expression during developmental stage transitions in *C. elegans*. let-7 was soon found to be conserved in many species, indicating the existence of a wider phenomenon.

MicroRNAs (miRNAs) are short ribonucleic acid (RNA) molecules, on average only 22 nucleotides long and are found in all eukaryotic cells. miRNAs are post-transcriptional regulators that bind to complementary sequences on target messenger RNA transcripts (mRNAs), usually resulting in translational repression and gene silencing.

The function of miRNAs appears to be in gene regulation. For that purpose, a miRNA is complementary to a part of one or more messenger RNAs (mRNAs). Animal miRNAs are usually complementary to a site in the 3' UTR whereas plant miRNAs are usually complementary to coding regions of mRNAs. Perfect or near perfect base pairing with the target RNA promotes cleavage of the RNA. This is the primary mode of plant microRNAs. In animals, microRNAs more often only partially base pair and inhibit protein translation of the target mRNA (this exists in plants as well but is less common). MicroRNAs that are partially complementary to a target can also speed up deadenylation, causing mRNAs to be degraded sooner. For partially complementary microRNAs to recognise their targets, nucleotides 2–7 of the miRNA (its ‘seed region’) still have to be perfectly complementary.^[68] miRNAs occasionally also cause histone modification and DNA methylation of promoter sites, which affects the expression of target genes. Animal microRNAs target in particular developmental genes. In contrast, genes involved in functions common to all cells, such as gene expression, have very few microRNA target sites and seem to be under selection to avoid targeting by microRNAs. dsRNA can also activate gene expression, a mechanism that has been termed “small RNA-induced gene activation” or RNAa. dsRNAs targeting gene promoters can induce potent transcriptional activation of associated genes. This was demonstrated in human cells using synthetic dsRNAs termed small

activating RNAs (saRNAs), but has also been demonstrated for endogenous microRNA.^[48]

miRNA and disease

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

miRNA and cancer

Several miRNAs have been found to have links with some types of cancer. A study of mice altered to produce excess c-Myc – a protein with mutated forms implicated in several cancers – shows that miRNA has an effect on the development of cancer. Mice that were engineered to produce a surplus of types of miRNA found in lymphoma cells developed the disease within 50 days and died two weeks later. In contrast, mice without the surplus miRNA lived over 100 days. Leukemia can be caused by the insertion of a viral genome next to the 17-92 array of microRNAs leading to increased expression of this microRNA. Another study found that two types of miRNA inhibit the E2F1 protein, which regulates cell proliferation. miRNA appears to bind to messenger RNA before it can be translated to proteins that switch genes on and off. By measuring activity among 217 genes encoding miRNA, patterns of gene activity that can distinguish types of cancers can be discerned. miRNA signatures may enable classification of cancer. This will allow doctors to determine the original tissue type which spawned a cancer and to be able to target a treatment course based on the original tissue type. miRNA profiling has already been able to determine whether patients with chronic lymphocytic leukemia had slow growing or aggressive forms of the cancer. Transgenic mice that over-express or lack specific miRNAs have provided insight into the role of small RNAs in various malignancies. A novel miRNA-profiling based screening assay for the detection of early-stage colorectal cancer has been developed and is currently in clinical trials. Early results showed that blood plasma samples collected

from patients with early, resectable (Stage II) colorectal cancer could be distinguished from those of sex-and age-matched healthy volunteers. Sufficient selectivity and specificity could be achieved using small (less than 1 mL) samples of blood. The test has potential to be a cost-effective, non-invasive way to identify at-risk patients who should undergo colonoscopy.

miRNA and heart disease

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

miRNA and the nervous system

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

Facts to be remembered

This is a **list of RNAs** in nature. Some of these categories are broad, others are single RNA families.

RNAs involved in protein synthesis

Type	Abbr.	Function	Distribution	Ref.
Messenger RNA	mRNA	Codes for protein	All organisms	
Ribosomal RNA	rRNA	Translation	All organisms	
Signal recognition particle RNA	7SL RNA or SRP RNA	Membrane integration	All organisms	[49]
Transfer RNA	tRNA	Translation	All organisms	
Transfer-messenger RNA	tmRNA	Rescuing stalled ribosomes	Bacteria	[50]

RNAs involved in post-transcriptional modification or DNA replication

Type	Abbr.	Function	Distribution	Ref.
Small nuclear RNA	snRNA	Splicing and other functions	Eukaryotes and archaea	[51]
Small nucleolar RNA	snoRNA	Nucleotide modification of RNAs	Eukaryotes and archaea	[52]
SmY RNA	SmY	mRNA trans-splicing	Nematodes	[53]
Small Cajal body-specific RNA	scaRNA	Type of snoRNA; Nucleotide modification of RNAs		
Guide RNA	gRNA	mRNA nucleotide modification	Kinetoplastid mitochondria	[54]
Ribonuclease P	RNase P	tRNA maturation	All organisms	[55]
Ribonuclease MRP	RNase MRP	rRNA maturation, DNA replication	Eukaryotes	[56]
Y RNA		RNA processing, DNA replication	Animals	[57]
Telomerase RNA		Telomere synthesis	Most eukaryotes	[58]

Regulatory RNAs

Type	Abbr.	Function	Distribution	Ref.
Antisense RNA	aRNA	Transcriptional attenuation / mRNA degradation / mRNA stabilisation / Translation block	All organisms	[59][60]
Cis-natural antisense transcript		Gene regulation		
CRISPR RNA	crRNA	Resistance to parasites, probably by targeting their DNA	Bacteria and archaea	[61]
Long noncoding RNA	Long ncRNA	Various	Eukaryotes	
MicroRNA	miRNA	Gene regulation	Most eukaryotes	[62]
Piwi-interacting RNA	piRNA	Transposon defense, maybe other functions	Most animals	[63][64]
Small interfering RNA	siRNA	Gene regulation	Most eukaryotes	[65]
Trans-acting siRNA	tasiRNA	Gene regulation	Land plants	[66]
Repeat associated siRNA	rasiRNA	Type of piRNA; transposon defense	Drosophila	[67]
7SK RNA	7SK	negatively regulating CDK9/cyclin T complex		

Parasitic RNAs

Type	Function	Distribution	Ref.
Retrotransposon	Self-propagating	Eukaryotes and some bacteria	[68]
Viral genome	Information carrier	Double-stranded RNA viruses, positive-sense RNA viruses, negative-sense RNA viruses, many satellite viruses and reverse transcribing viruses	
Viroid	Self-propagating	Infected plants	[69]
Satellite RNA	Self-propagating	Infected cells	

Other RNAs

Type	Abbr.	Function	Distribution	Ref.
Vault RNA	vRNA	Expulsion of xenobiotics, maybe		[70]

References

1. ↑ RNA
2. ↑ Dahm R (2005). “Friedrich Miescher and the discovery of DNA”. *Developmental Biology* **278** (2): 274–88. doi:10.1016/j.ydbio.2004.11.028. PMID 15680349.
3. ↑ Caspersson T, Schultz J (1939). “Pentose nucleotides in the cytoplasm of growing tissues”. *Nature* **143**: 602–3. doi:10.1038/143602c0.
4. ↑ Ochoa S (1959). “Enzymatic synthesis of ribonucleic acid”. *Nobel Lecture*. http://nobelprize.org/nobel_prizes/medicine/laureates/1959/ochoa-lecture.pdf.
5. ↑ Holley RW et al. (1965). “Structure of a ribonucleic acid”. *Science* **147** (1664): 1462–65. doi:10.1126/science.147.3664.1462. PMID 14263761.

6. ↑ Siebert S (2006). "Common sequence structure properties and stable regions in RNA secondary structures". Dissertation, Albert-Ludwigs-Universität, Freiburg im Breisgau. pp. 1. http://deposit.ddb.de/cgi-bin/dokserv?idn=982323891&dok_var=d1&dok_ext=pdf&filename=982323891.pdf.
7. ↑ Szathmáry E (1999). "The origin of the genetic code: amino acids as cofactors in an RNA world". *Trends Genet.* **15** (6): 223–9. doi:10.1016/S0168-9525(99)01730-8. PMID 10354582.
8. ↑ Fiers W et al. (1976). "Complete nucleotide-sequence of bacteriophage MS2-RNA: primary and secondary structure of replicase gene". *Nature* **260** (5551): 500–7. doi:10.1038/260500a0. PMID 1264203.
9. ↑ Napoli C, Lemieux C, Jorgensen R (1990). "Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans". *Plant Cell* **2** (4): 279–89. doi:10.1105/tpc.2.4.279. PMID 12354959.
10. ↑ Dafny-Yelin M, Chung SM, Frankman EL, Tzfira T (December 2007). "pSAT RNA interference vectors: a modular series for multiple gene down-regulation in plants". *Plant Physiol.* **145** (4): 1272–81. doi:10.1104/pp.107.106062. PMID 17766396.
11. ↑ Ruvkun G (2001). "Glimpses of a tiny RNA world". *Science* **294** (5543): 797–99. doi:10.1126/science.1066315. PMID 11679654.
12. ↑ Fichou Y, Férec C (2006). "The potential of oligonucleotides for therapeutic applications". *Trends in Biotechnology* **24** (12): 563–70. doi:10.1016/j.tibtech.2006.10.003. PMID 17045686.
13. ↑ Small nucleolar RNA
14. ↑ RNA
15. ↑ <http://www.madsci.org>
16. ↑ Horton, Robert H.; et al. *Principles of Biochemistry*. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002.
17. ↑ Zorbach, W.W. *Synthetic Procedures in Nucleic Acid Chemistry: Physical and Physicochemical Aids in Determination of Structure*. Vol 2. New York: Wiley-

- Interscience, 1973.
- 18. ↑ Lee, J.K.; Kurinovich, Ma. *J Am Soc Mass Spectrom.* **13**(8), 2005, 985–95.
 - 19. ↑ SnRNP
 - 20. ↑ Enzyme definition Dictionary.com Accessed 6 April 2007
 - 21. ↑ Carl Woese, *The Genetic Code* (New York: Harper and Row, 1967).
 - 22. ↑ The Nobel Prize in Chemistry 1989 was awarded to Thomas R. Cech and Sidney Altman “for their discovery of catalytic properties of RNA”.
 - 23. ↑ Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR (November 1982). “Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena”. *Cell* **31** (1): 147–57. PMID 6297745.
 - 24. ↑ Forster AC, Symons RH (1987). “Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites”. *Cell* **49** (2): 211–220. doi:10.1016/0092-8674(87)90562-9. PMID 2436805.
 - 25. ↑ Johnston W, Unrau P, Lawrence M, Glasner M, Bartel D (2001). “RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension” (PDF). *Science* **292** (5520): 1319–25. doi:10.1126/science.1060786. PMID 11358999. http://web.wi.mit.edu/bartel/pub/publication_reprints/Johnston_Science01.pdf.
 - 26. ↑ W. K. Johnston, P. J. Unrau, M. S. Lawrence, M. E. Glasner and D. P. Bartel RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension. *Science* **292**, 1319 (2001)
 - 27. ↑ Hani S. Zaher and Peter J. Unrau, Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* (2007), 13:1017-1026
 - 28. ↑ Huang, Yang, and Yarus, RNA enzymes with two small-molecule substrates. *Chemistry & Biology*, Vol 5, 669–678, November 1998
 - 29. ↑ Unrau, P. J.; Bartel, D. P. (1998). “RNA-catalysed nucleotide

- synthesis". *Nature* **395** (6699): 260–263. doi:10.1038/26193. PMID 9751052.
30. ↑ Zhang, Biliang; Cech, Thomas R. (1997). "Peptide bond formation by *in vitro* selected ribozymes". *Nature* **390** (6655): 96–100. doi:10.1038/36375. PMID 9363898.
 31. ↑ Szathmary, E. (1999). "The origin of the genetic code: amino acids as cofactors in an RNA world". *Trends in Genetics* **15** (6): 223–229. doi:10.1016/S0168-9525(99)01730-8. PMID 10354582.
 32. ↑ Ribonuclease
 33. ↑ http://en.wikipedia.org/w/index.php?title=Ribonuclease_A&oldid=417169401
 34. ↑ http://en.wikipedia.org/w/index.php?title=RNase_H&oldid=422163070
 35. ↑ Macrae I, Zhou K, Li F, Repic A, Brooks A, Cande W, Adams P, Doudna J (2006). "Structural basis for double-stranded RNA processing by dicer". *Science* **311** (5758): 195–8. doi:10.1126/science.1121638. PMID 16410517.
 36. ↑ Ecker JR, Davis RW (1986). "Inhibition of gene expression in plant cells by expression of antisense RNA". *Proc Natl Acad Sci USA* **83** (15): 5372–5376. doi:10.1073/pnas.83.15.5372. PMID 16593734.
 37. ↑ Napoli C, Lemieux C, Jorgensen R (1990). "Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in *trans*". *Plant Cell* **2** (4): 279–289. doi:10.1105/tpc.2.4.279. PMID 12354959.
 38. ↑ Romano N, Macino G (1992). "Quelling: transient inactivation of gene expression in *Neurospora crassa* by transformation with homologous sequences". *Mol Microbiol* **6** (22): 3343–53. doi:10.1111/j.1365-2958.1992.tb02202.x. PMID 1484489.
 39. ↑ Van Blokland R, Van der Geest N, Mol JNM, Kooter JM (1994). "Transgene-mediated suppression of chalcone synthase expression in *Petunia hybrida* results from an increase in RNA turnover". *Plant J* **6**: 861–77. doi:10.1046/j.1365-313X.1994.6060861.x/abs/. <http://www.blackwell-journals.com>

- synergy.com/links/doi/10.1046/j.1365-313X.1994.6060861.x/abs/.
40. ↑ Mol JNM, van der Krol AR (1991). *Antisense nucleic acids and proteins: fundamentals and applications*. M. Dekker. pp. 4, 136. ISBN 0824785169.
 41. ↑ Matzke MA, Matzke AJM. (2004). “Planting the Seeds of a New Paradigm”. *PLoS Biol* **2** (5): e133. doi:10.1371/journal.pbio.0020133. PMID 15138502.
 42. ↑ Covey S, Al-Kaff N, Lángara A, Turner D (1997). “Plants combat infection by gene silencing”. *Nature* **385**: 781–2. doi:10.1038/385781a0.
 43. ↑ Ratcliff F, Harrison B, Baulcombe D (1997). “A Similarity Between Viral Defense and Gene Silencing in Plants”. *Science* **276**: 1558–60. doi:10.1126/science.276.5318.1558. PMID 18610513.
 44. ↑ Guo S, Kemphues K (1995). “par-1, a gene required for establishing polarity in *C. elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed”. *Cell* **81** (4): 611–20. doi:10.1016/0092-8674(95)90082-9. PMID 7758115.
 45. ↑ Pal-Bhadra M, Bhadra U, Birchler J (1997). “Cosuppression in *Drosophila*: gene silencing of Alcohol dehydrogenase by white-Adh transgenes is Polycomb dependent”. *Cell* **90** (3): 479–90. doi:10.1016/S0092-8674(00)80508-5. PMID 9267028.
 46. ↑ Bagasra O, Prilliman KR (2004). “RNA interference: the molecular immune system”. *J. Mol. Histol.* **35** (6): 545–53. doi:10.1007/s10735-004-2192-8. PMID 15614608.
[http://www.kluweronline.com/
art.pdf?issn=1567-2379&volume=35&page=545](http://www.kluweronline.com/art.pdf?issn=1567-2379&volume=35&page=545).
 47. ↑ Jaronczyk K, Carmichael J, Hobman T (2005). “Exploring the functions of RNA interference pathway proteins: some functions are more RISCy than others?”. *Biochem J* **387** (Pt 3): 561–71. doi:10.1042/BJ20041822. PMID 15845026. PMC 1134985.
[http://www.ncbi.nlm.nih.gov/
pmc/articles/PMC1134985/articlerender.fcgi?tool=pubmed&pubmedid=15845026](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1134985/articlerender.fcgi?tool=pubmed&pubmedid=15845026).
 48. ↑ MicroRNA

49. ↑ Gribaldo S, Brochier-Armanet C (2006). “The origin and evolution of Archaea: a state of the art”. *Philos Trans R Soc Lond B Biol Sci.* **361** (1470): 1007–22. doi:10.1098/rstb.2006.1841. PMID 16754611.
50. ↑ Gillet R, Felden B (2001). “Emerging views on tmRNA-mediated protein tagging and ribosome rescue”. *Molecular Microbiology* **42** (4): 879–85. doi:10.1046/j.1365-2958.2001.02701.x. PMID 11737633.
51. ↑ Thore S, Mayer C, Sauter C, Weeks S, Suck D (2003). “Crystal Structures of the Pyrococcus abyssi Sm Core and Its Complex with RNA”. *J. Biol. Chem.* **278** (2): 1239–47. doi:10.1074/jbc.M207685200. PMID 12409299.
52. ↑ Kiss T (2001). “Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs”. *The EMBO Journal* **20** (14): 3617–22. doi:10.1093/emboj/20.14.3617. PMID 11447102.
53. ↑ Jones TA, Otto W, Marz M, Eddy SR, Stadler PF (2009). “A survey of nematode SmY RNAs”. *RNA Biol* **6** (1): 5–8. doi:10.4161/rna.6.1.7634. PMID 19106623.
<http://www.landesbioscience.com/journals/rna/abstract.php?id=7634>.
54. ↑ Alfonzo JD, Thiemann O, Simpson L (1997). “The mechanism of U insertion/deletion RNA editing in kinetoplastid mitochondria”. *Nucleic Acids Research* **25** (19): 3751–59. doi:10.1093/nar/25.19.3751. PMID 9380494.
55. ↑ Pannucci JA, Haas ES, Hall TA, Harris JK, Brown JW (1999). “RNase P RNAs from some Archaea are catalytically active”. *Proc Natl Acad Sci USA* **96** (14): 7803–08. doi:10.1073/pnas.96.14.7803. PMID 10393902.
56. ↑ Woodhams MD, Stadler PF, Penny D, Collins LJ (2007). “RNase MRP and the RNA processing cascade in the eukaryotic ancestor”. *BMC Evolutionary Biology* **7**: S13. doi:10.1186/1471-2148-7-S1-S13. PMID 17288571.
57. ↑ Perreault J, Perreault J-P, Boire G (2007). “Ro-associated Y RNAs in metazoans: evolution and diversification”. *Molecular*

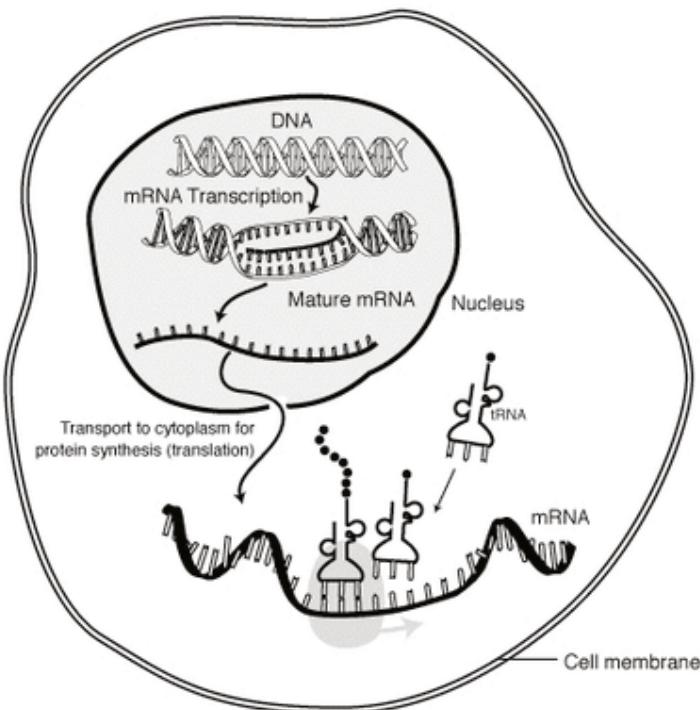
- Biology and Evolution* **24** (8): 1678–89. doi:10.1093/molbev/msm084. PMID 17470436.
- 58. ↑ Lustig AJ (1999). “Crisis intervention: The role of telomerase”. *Proc Natl Acad Sci USA* **96** (7): 3339–41. doi:10.1073/pnas.96.7.3339. PMID 10097039.
 - 59. ↑ Brantl S (2002). “Antisense-RNA regulation and RNA interference”. *Biochimica et Biophysica Acta* **1575** (1–3): 15–25. PMID 12020814.
 - 60. ↑ Brantl S (2007). “Regulatory mechanisms employed by cis-encoded antisense RNAs”. *Curr. Opin. Microbiol.* **10** (2): 102–9. doi:10.1016/j.mib.2007.03.012. PMID 17387036.
 - 61. ↑ Brouns SJ, Jore MM, Lundgren M, et al. (August 2008). “Small CRISPR RNAs guide antiviral defense in prokaryotes”. *Science* **321** (5891): 960–4. doi:10.1126/science.1159689. PMID 18703739.
 - 62. ↑ Lin S-L, Miller JD, Ying S-Y (2006). “Intronic microRNA (miRNA)”. *Journal of Biomedicine and Biotechnology* **2006** (4): 1–13. doi:10.1155/JBB/2006/26818. PMID 17057362.
 - 63. ↑ Horwich MD, Li C Matranga C, Vagin V, Farley G, Wang P, Zamore PD (2007). “The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC”. *Current Biology* **17** (14): 1265–72. doi:10.1016/j.cub.2007.06.030. PMID 17604629.
 - 64. ↑ Ghildiyal M, Zamore PD (February 2009). “Small silencing RNAs: an expanding universe”. *Nat. Rev. Genet.* **10** (2): 94–108. doi:10.1038/nrg2504. PMID 19148191.
 - 65. ↑ Ahmad K, Henikoff S (2002). “Epigenetic consequences of nucleosome dynamics”. *Cell* **111** (3): 281–84. doi:10.1016/S0092-8674(02)01081-4.
 - 66. ↑ Vazquez F, Vaucheret H (2004). “Endogenous trans-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs”. *Mol. Cell* **2006** (16): 1–13. doi:10.1155/JBB/2006/26818. PMID 17057362.
 - 67. ↑ Dasset S, Buchon N, Meignin C, Coiffet M, Vaury C (2008). “In *Drosophila melanogaster* the COM locus directs the somatic silencing of two retrotransposons through both Piwi-

- dependent and -independent pathways". PLoS ONE **3** (2): e1526. doi:10.1371/journal.pone.0001526. PMID 18253480.
- 68. ↑ Boeke JD (2003). "The unusual phylogenetic distribution of retrotransposons: a hypothesis". *Genome Research* **13** (9): 1975–83. doi:10.1101/gr.1392003. PMID 12952870.
 - 69. ↑ Flores R, Hernández C, Martínez de Alba AE, Daròs JA, Di Serio F (2005). "Viroids and viroid-host interactions". *Annual Review of Phytopathology* **43**: 117–39. doi:10.1146/annurev.phyto.43.040204.140243. PMID 16078879.
 - 70. ↑ Gopinath SC, Matsugami A, Katahira M, Kumar PK (2005). "Human vault-associated non-coding RNAs bind to mitoxantrone, a chemotherapeutic compound". *Nucleic Acids Res.* **33** (15): 4874–81. doi:10.1093/nar/gki809. PMID 16150923.

I5.

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

In transcription, the codons of a gene are copied into messenger RNA by RNA polymerase. In simple word or formation of RNA from DNA is known as transcription. 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 (4³ 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.



Contents

- 1 Transcription
- 2 Formation of Pre-initiation complex
 - 2.1 TATA-binding protein and Transcription factor II D
- 3 RNA polymerase
- 4 RNA polymerase in eukaryotes
 - 4.1 RNA polymerase I
 - 4.2 RNA polymerase II
 - 4.2.1 C-terminal domain (CTD) of RNA Pol II
 - 4.3 RNA polymerase III

- 5 RNA polymerase IV
- 6 Initiation of transcription
 - 6.1 Transcription Factors
- 7 Elongation of RNA
- 8 Transcription termination
 - 8.1 Rho-dependent termination
 - 8.2 Rho-independent termination
- 9 mRNA and its modification
 - 9.1 Capping of mRNA
 - 9.2 Polyadenylation of mRNA
- 10 Splicing of RNA
 - 10.1 Spliceosome and its assembly
- 11 Self splicing
- 12 RNA-editing
 - 12.1 Editing by insertion or deletion
- 13 References

Transcription

Transcription can be explained easily in 4 or 5 simple steps, each moving like a wave along the DNA.

Unwinding of DNA /"unzips" as the Hydrogen Bonds Break.

The free nucleotides of the RNA, pair with complementary DNA bases.

RNA sugar-phosphate backbone forms. (Aided by RNA Polymerase.)

Hydrogen bonds of the untwisted RNA+DNA "ladder" break, freeing the new RNA.

If the cell has a nucleus, the RNA is further processed and then moves through the small nuclear pores to the cytoplasm.

The major steps of transcription process are:

1. Formation of Pre-initiation complex
2. Initiation of transcription
3. Promoter clearance
4. Elongation of RNA
5. Termination

Formation of Pre-initiation complex

In eukaryotes, RNA polymerase, and therefore the initiation of transcription, requires the presence of a core **promoter** sequence in the DNA. Promoters are regions of DNA that promote transcription and, in eukaryotes, are found at -30, -75, and -90 base pairs upstream from the start site of transcription. Core promoters are sequences within the promoter that are essential for transcription initiation. RNA polymerase is able to bind to core promoters in the presence of various specific transcription factors.

The most common type of core promoter in eukaryotes is a short DNA sequence known as a TATA box, found 25-30 base pairs upstream from the start site of transcription. The TATA box, as a core promoter, is the binding site for a transcription factor known as TATA-binding protein (TBP), which is itself a subunit of another transcription factor, called Transcription Factor II D (TFIID). After TFIID binds to the TATA box via the TBP, five more transcription factors and RNA polymerase combine around the TATA box in a series of stages to form a preinitiation complex. One transcription factor, DNA helicase, has helicase activity and so is involved in the separating of opposing strands of double-stranded DNA to provide access to a single-stranded DNA template. However, only a low, or basal, rate of transcription is driven by the preinitiation complex alone. Other proteins known as activators and repressors, along with any associated coactivators or corepressors, are responsible for modulating transcription rate.

Thus, preinitiation complex contains:

- Core Promoter Sequence,
- Transcription Factors,
- DNA Helicase,
- RNA Polymerase,
- Activators and Repressors.

The transcription preinitiation in archaea is, in essence, homologous to that of eukaryotes, but is much less complex.^[1] The archaeal preinitiation complex assembles at a TATA-box binding site; however, in archaea, this complex is composed of only RNA polymerase II, TBP, and TFB (the archaeal homologue of eukaryotic transcription factor II B (TFIIB)).^{[2][3]} Typically the PIC is made up of six general transcription factors: TFIIA (GTF2A1, GTF2A2), TFIIB (GTF2B), B-TFIID (BTAF1, TBP), TFIID (BTAF1, BTF3, BTF3L4, EDF1, TAF1-15, 16 total), TFIIE, TFIIF, and TFIIH. Also, at some point during its assembly it is joined with RNA polymerase II and the remaining components of the holoenzyme.

1. The TATA binding protein (TBP, a subunit of TFIID), TBPL1, or TBPL2 can bind the promoter or TATA box. Most genes lack a TATA box and use an initiator element (INR) or downstream core promoter instead. Nevertheless, TBP is always involved and is forced to bind without sequence specificity. TAFs from TFIID can also be involved when the TATA box is absent. A TFIID TAF will bind sequence specifically, and force the TBP to bind non-sequence specifically, bringing the remaining portions of TFIID to the promoter.
2. TFIIA interacts with the TBP subunit of TFIID and aids in the binding of TBP to TATA-box containing promoter DNA. Although TFIIA does not recognize DNA itself, its interactions with TBP allow it to stabilize and facilitate formation of the PIC.
3. The N-terminal domain of TFIIB brings the DNA into proper position for entry into the active site of RNA polymerase II. TFIIB binds partially sequence specifically, with some

preference for BRE. The TFIID-TFIIA-TFIIB (DAB)-promoter complex subsequently recruits RNA polymerase II and TFIIF.

4. TFIIF (two subunits, RAP30 and RAP74, showing some similarity to bacterial sigma factors) and Pol II enter the complex together. TFIIF helps to speed up the polymerization process.
5. TFIIE joins the growing complex and recruits TFIIH. TFIIE may be involved in DNA melting at the promoter: it contains a zinc ribbon motif that can bind single stranded DNA. TFIIE helps to open and close the Pol II's 'Jaw' like structure, which enables movement down the DNA strand.
6. DNA may be wrapped one complete turn around the preinitiation complex and it is TFIIF that helps keep this tight wrapping. In the process, the torsional strain on the DNA may aid in DNA melting at the promoter, forming the transcription bubble.
7. TFIIH and TFIIJ enter the complex together. TFIIH is a large protein complex that contains among others the CDK7/cyclin H kinase complex and a DNA helicase. TFIIH has three functions: it binds specifically to the template strand to ensure that the correct strand of DNA is transcribed and melts or unwinds the DNA (ATP dependently) to separate the two strands using its helicase activity. It has a kinase activity that phosphorylates the C-terminal domain (CTD) of Pol II at the amino acid serine. This switches the RNA polymerase to start producing RNA. Finally it is essential for Nucleotide Excision Repair (NER) of damaged DNA. TFIIH and TFIIE strongly interact with one another. TFIIE affects TFIIH's catalytic activity. Without TFIIE, TFIIH will not unwind the promoter.
8. TFIIH helps create the transcription bubble and may be required for transcription if the DNA template is not already denatured or if it is supercoiled.
9. Mediator then encases all the transcription factors and Pol II. It interacts with enhancers, areas very far away (upstream or downstream) that help regulate transcription.

The formation of the transcription preinitiation complex (PIC) is analogous to the mechanism seen in bacterial initiation. In bacteria, the sigma factor recognizes and binds to the promoter sequence. In eukaryotes, the transcription factors perform this role.

TATA-binding protein and Transcription factor II D

TBP is a subunit of the eukaryotic transcription factor **TFIID**. TFIID is the first protein to bind to DNA during the formation of the pre-initiation transcription complex of RNA polymerase II (RNA Pol II). Binding of TFIID to the TATA box in the promoter region of the gene initiates the recruitment of other factors required for RNA Pol II to begin transcription. Some of the other recruited transcription factors include TFIIA, TFIIB, and TFIIF. Each of these transcription factors is formed from the interaction of many protein subunits, indicating that transcription is a heavily regulated process. TBP is also a necessary component of RNA polymerase I and RNA polymerase III, and is, it is thought, the only common subunit required by all three of the RNA polymerases.^[4]

The TATA-binding protein (TBP) is a transcription factor that binds specifically to a DNA sequence called the TATA box. This DNA sequence is found about 35 base pairs upstream of the transcription start site in some eukaryotic gene promoters. TBP, along with a variety of TBP-associated factors, make up the TFIID, a general transcription factor that in turn makes up part of the RNA polymerase II preinitiation complex. As one of the few proteins in the preinitiation complex that binds DNA in a sequence-specific manner, it helps position RNA polymerase II over the transcription start site of the gene. However, it is estimated that only 10–20% of human promoters have TATA boxes. Therefore, TBP is probably not the only protein involved in positioning RNA polymerase II.

Transcription factor II D (TFIID) is one of several general

transcription factors that make up the RNA polymerase II preinitiation complex. Before the start of transcription, the transcription Factor II D (TFIID) complex, consisting of TFIID, TBP, and at least nine other polypeptides, binds to the TATA box in the core promoter of the gene. TFIID is itself composed of several subunits called TBP-associated factors (TAFs, of which there are 16) and the TATA Binding Protein (TBP). In a test tube, only TBP is necessary for transcription at promoters that contain a TATA box. TAFs, however, add promoter selectivity, especially if there is no TATA box sequence for TBP to bind to. TAFs are included in two distinct complexes, TFIID and B-TFIID. The TFIID complex is composed of TBP and more than eight TAFs. But, the majority of TBP is present in the B-TFIID complex, which is composed of TBP and TAFII170 (BTAF1) in a 1:1 ratio.

RNA polymerase

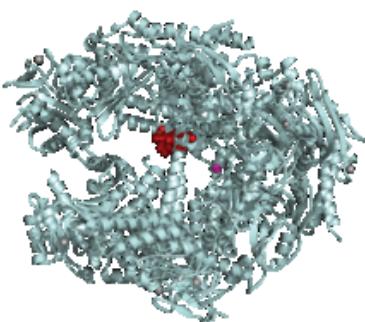
RNAP was discovered independently by Sam Weiss, Audrey Stevens, and Jerard Hurwitz in 1960.^[5]

RNA polymerase (RNAP) is an enzyme that produces RNA. In cells, RNAP is needed for constructing RNA chains from DNA genes as templates, a process called transcription. RNA polymerase enzymes are essential to life and are found in all organisms and many viruses. In chemical terms, RNAP is a nucleotidyl transferase that polymerizes ribonucleotides at the 3' end of an RNA transcript.

RNA polymerase in eukaryotes

Eukaryotes have several types of RNAP, characterized by the type of RNA they synthesize: The following RNA polymerases are very common in eukaryotic cells.

- 1 RNA polymerase I
- 2 RNA polymerase II
- 3 RNA polymerase III
- 4 RNA polymerase IV
- 5 RNA polymerase V



Structure of eukaryotic RNA polymerase II (light blue) in complex with α -amanitin (red), a strong poison found in death cap mushrooms that targets this vital enzyme

RNA polymerase I

RNA polymerase I synthesizes a pre-rRNA 45S, which matures into 28S, 18S and 5.8S rRNAs which will form the major RNA sections of the ribosome.^[6] Pol I consists of 8-14 protein subunits (polypeptides). All 12 subunits have identical or related counterparts in PolII and Pol III. rDNA transcription is confined to the nucleolus where several hundreds of copies of rRNA genes are present, arranged as tandem head-to-tail repeats.

Pol I transcribes one large transcript, encoding an rDNA gene over and over again. This gene encodes the 18S, the 5.8S, and the 28S RNA molecules of the ribosome in eukaryotes. The transcripts are cleaved by snoRNA. The 5S ribosomal RNA is transcribed by Pol III. Because of the simplicity of Pol I transcription, it is the fastest-acting polymerase. When rRNA synthesis is stimulated, SL1 (selectivity factor 1) will bind to the promoters of rDNA genes that were previously silent, and recruit a pre-initiation complex to which Pol I will bind and start transcription of rRNA. Changes in rRNA

transcription can also occur via changes in the rate of transcription. While the exact mechanism through which Pol I increases its rate of transcription is yet unknown, evidence has shown that rRNA synthesis can increase or decrease without changes in the number of actively transcribed rDNA.

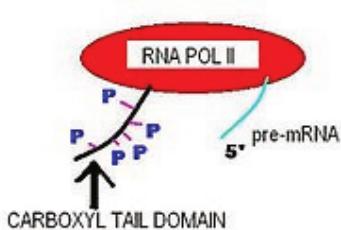
RNA polymerase II



RNA polymerase II synthesizes precursors of mRNAs and most snRNA and microRNAs.^[7] This is the most studied type, and due to the high level of control required over transcription a range of transcription factors are required for its binding to promoters. RNA polymerase II is a 550 kDa complex which contains 12 subunits. The eukaryote|eukaryotic core RNA polymerase II was first purified using transcription assays.^[8] The purified enzyme has typically 10-12 subunits (12 in humans and yeast) and is incapable of specific promoter recognition.^[9] Many subunit-subunit interactions are known.^[10] Computer generated image of POLR2A gene colorized subunits: **green** – RPB1 domain 1, **blue** – RPB1 domain 2, **sand** – RPB1 domain 3, **light blue** – RPB1 domain 4, **brown** – RPB1 domain 6, and **magenta**– RPB1 CTD.DNA-directed RNA polymerase II subunit

RPB1 is an enzyme that in humans is encoded by the POLR2A gene. RPB1 is the largest subunit of RNA polymerase II. It contains a C-terminus|carboxy terminal domain (CTD) composed of up to 52 heptapeptide repeats (YSPTSPS) that are essential for polymerase activity.^[11] In combination with several other polymerase subunits, it forms the DNA binding domain of the polymerase, a groove in which the DNA template is transcribed into RNA.^[12] It strongly interacts with RPB8.^[10] RPB2 (POLR2B) is the second largest subunit which in combination with at least two other polymerase subunits forms a structure within the polymerase that maintains contact in the active site of the enzyme between the DNA template and the newly synthesized RNA.^[13] The third largest subunit RPB3 (POLR2C) exists as a heterodimer with POLR2J forming a core subassembly. RPB3 strongly interacts with RPB1-5, 7, 10-12.^[10] RNA polymerase II subunit B4 (RPB4) encoded by the POLR2D gene^[14] is the fourth largest subunit and may have a stress protective role. In humans RPB5 is encoded by the POLR2E gene. Two molecules of this subunit are present in each RNA polymerase II.^[15] RPB5 strongly interacts with RPB1, RPB3, and RPB6.^[10] RPB6 (POLR2F) forms a structure with at least two other subunits that stabilizes the transcribing polymerase on the DNA template.^[16] POLR2G encodes RPB7 that may play a role in regulating polymerase function.^[17] RPB7 interacts strongly with RPB1 and RPB5.^[10] RPB8 (POLR2H) interacts with subunits RPB1-3, 5, and 7.^[10] The groove in which the DNA template is transcribed into RNA is composed of RPB9 (POLR2I) and RPB1. RPB10 is the product of gene POLR2L. It interacts with RPB1-3 and 5, and strongly with RPB3.^[10] The RPB11 subunit is itself composed of three subunits in humans: POLR2J (RPB11-a), POLR2J2 (RPB11-b), and POLR2J3^[18] (RPB11-c). Also interacting with RPB3 is RPB12 (POLR2K).^[10]

C-terminal domain (CTD) of RNA Pol II



RNA Pol II in action, showing the CTD extension to the C-terminal of POLR2A.

RNAPII can exist in two forms: RNAPII0, with a highly phosphorylated CTD, and RNAPII_A, with a nonphosphorylated CTD. The carboxy-terminal domain (CTD) of RNA polymerase II is that portion of the polymerase that is involved in the initiation of transcription, the capping of

the RNA transcript, and attachment to the spliceosome for RNA splicing.^[19] The CTD typically consists of up to 52 repeats of the sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser.^[20] The carboxy-terminal repeat domain (CTD) is essential for life. Cells containing only RNAPII with none or only up to one-third of its repeats are inviable.^[21]

The CTD is an extension appended to the C terminus of RPB1, the largest subunit of RNA polymerase II. It serves as a flexible binding scaffold for numerous nuclear factors, determined by the phosphorylation patterns on the CTD repeats. Each repeat contains an evolutionary conserved and repeated heptapeptide, Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7, which is subjected to reversible phosphorylations during each transcription cycle.^[22] This domain is inherently unstructured yet evolutionarily conserved, and in eukaryotes it comprises from 25 to 52 tandem copies of the consensus repeat heptad.^[21] As the CTD is frequently not required for general transcription factor (GTF)-mediated initiation and RNA synthesis, it does not form a part of the catalytic essence of RNAPII, but performs other functions.^{[22][23]}

Phosphorylation occurs principally on Ser2 and Ser5 of the repeats of CTD, although these positions are not equivalent. The phosphorylation state changes as RNAPII progresses through the transcription cycle: The initiating RNAPII is form IIA, and the

elongating enzyme is form II0. While RNAPII0 does consist of RNAPs with hyperphosphorylated CTDs, the pattern of phosphorylation on individual CTDs can vary due to differential phosphorylation of Ser2 versus Ser5 residues and/or to differential phosphorylation of repeats along the length of the CTD. The PCTD (phosphoCTD of an RNAPII0) physically links pre-mRNA processing to transcription by tethering processing factors to elongating RNAPII, e.g., 5'-end capping, 3'-end cleavage, and polyadenylation. Ser5 phosphorylation (Ser5PO₄) near the 5' ends of genes depends principally on the kinase activity of TFIIH (Kin28 in yeast; CDK7 in metazoans). The transcription factor TFIIH is a kinase and will hyperphosphorylate the CTD of RNAP, and in doing so, causes the RNAP complex to move away from the initiation site. Subsequent to the action of TFIIH kinase, Ser2 residues are phosphorylated by CTDK-I in yeast (CDK9 kinase in metazoans). Ctk1 (CDK9) acts in compliment to phosphorylation of serine 5 and is, thus, seen in middle to late elongation. CDK8 and cyclin C (CCNC) are components of the RNA polymerase II holoenzyme that phosphorylate the carboxyl-terminal domain (CTD). CDK8 regulates transcription by targeting the CDK7/cyclin H subunits of the general transcription initiation factor IIH (TFIIH), thereby providing a link between the mediator and the basal transcription machinery. The gene CTDPI1 encodes a phosphatase that interacts with the carboxy-terminus of transcription initiation factor TFIIF, a transcription factor that regulates elongation as well as initiation by RNA polymerase II^[24].

RNA polymerase III

RNA polymerase III synthesizes tRNAs, rRNA 5S and other small RNAs found in the nucleus and cytosol.^[25] RNA polymerase III (also called Pol III) transcribes DNA to synthesize ribosomal 5S rRNA, tRNA and other small RNAs. The genes transcribed by RNA Pol

III fall in the category of “housekeeping” genes whose expression is required in all cell types and most environmental conditions. Therefore the regulation of Pol III transcription is primarily tied to the regulation of cell growth and the cell cycle, thus requiring fewer regulatory proteins than RNA polymerase II. In the process of transcription (by any polymerase) there are three main stages: Initiation; requiring construction of the RNA polymerase complex on the gene’s promoter. Elongation; the writing of the RNA transcript. Termination; the finishing of RNA writing and disassembly of the RNA polymerase complex.

RNA polymerase IV

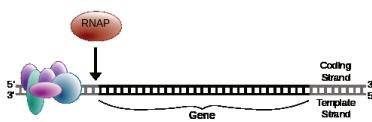
RNA polymerase IV synthesizes siRNA in plants.^[26] Polymerase IV is specific to plants genomes and is required for the synthesis of over 90% of all siRNA. RNA polymerase silences the transposons and repetitive DNA in the siRNA pathway. The siRNA plays a major role in defending the genome against the invading viruses and transposable elements by RNA directed DNA methylation. Polymerase IV and ROS1 demethylase unlocks and recondenses the 5S rDNA chromatin, which is present in seed and used for the development of adult features in plants. Polymerase IV is involved in setting the methylation patterns in the 5S genes during plant maturation. In *arabidopsis* polymerase IV works with binding protein DCL3 and a RNA polymerase II RDR2 in a silencing pathway which Polymerase IV would produce RNA, which is changed to dsRNA by RDR2 then converted to siRNA by DCL3.

- RNA polymerase V synthesizes RNAs involved in siRNA-directed heterochromatin formation in plants.^[27]

There are other RNA polymerase types in mitochondria and

chloroplasts. And there are RNA-dependent RNA polymerases involved in RNA interference.^[28]

Initiation of transcription



Simple diagram of transcription initiation. RNA polymerase (RNAP)

Transcription initiation is more complex in eukaryotes. Eukaryotic RNA polymerase does not directly recognize the core promoter sequences. Instead, a collection of proteins

called transcription factors mediate the binding of RNA polymerase and the initiation of transcription. Only after certain transcription factors are attached to the promoter does the RNA polymerase bind to it. The completed assembly of transcription factors and RNA polymerase bind to the promoter, forming a transcription initiation complex. Transcription in the archaea domain is similar to transcription in eukaryotes.^[29]

In bacteria, transcription begins with the binding of RNA polymerase to the promoter in DNA. RNA polymerase is a core enzyme consisting of five subunits: 2 α subunits, 1 β subunit, 1 β' subunit, and 1 ω subunit. At the start of initiation, the core enzyme is associated with a sigma factor that aids in finding the appropriate -35 and -10 base pairs downstream of promoter sequences.

What is sigma factor?

A sigma factor (σ factor) is a prokaryotic transcription initiation factor that enables specific binding of RNA polymerase to gene promoters. Different sigma factors are activated in response to different environmental conditions. Every molecule of RNA polymerase contains exactly one sigma factor subunit, which in the model bacterium *Escherichia coli* is one of those listed below. *E. coli* has seven sigma factors; the number of sigma factors varies between bacterial species. Sigma factors are distinguished by their

characteristic molecular weights. For example, σ 70 refers to the sigma factor with a molecular weight of 70 kDa.

Transcription Factors

Transcription factors are essential for the regulation of gene expression and are, as a consequence, found in all living organisms. The number of transcription factors found within an organism increases with genome size, and larger genomes tend to have more transcription factors per gene. There are approximately 2600 proteins in the human genome that contain DNA-binding domains, and most of these are presumed to function as transcription factors. Therefore, approximately 10% of genes in the genome code for transcription factors, which makes this family the single largest family of human proteins. Furthermore, genes are often flanked by several binding sites for distinct transcription factors, and efficient expression of each of these genes requires the cooperative action of several different transcription factors (see, for example, hepatocyte nuclear factors). Hence, the combinatorial use of a subset of the approximately 2000 human transcription factors easily accounts for the unique regulation of each gene in the human genome during development.

In molecular biology, a transcription factor (sometimes called a sequence-specific DNA-binding factor) is a protein that binds to specific DNA sequences, thereby controlling the movement (or transcription) of genetic information from DNA to mRNA. Transcription factors perform this function alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes.

General transcription factors or GTFs are intimately involved in the process of gene regulation, and most are required for life. TATA

binding protein, (TBP) is a GTF that binds to the TATAA box (T=Thymine, A=Adenine) the motif of nucleic acids that is directly upstream from the coding region in all genes. TBP is responsible for the recruitment of the RNA Pol II holoenzyme, the final event in transcription initiation. These proteins are ubiquitous and interact with the core promoter region of DNA, which contains the transcription start site(s) of all class II genes. Not all GTFs play a role in transcriptional initiation; some are required for the second general step in transcription, elongation. For example, members of the FACT complex (Spt16/Pob3 in *S. cerevisiae*, SUPT16H/SSRP1 in humans) facilitate the rapid movement of RNA Pol II over the encoding region of genes. This is accomplished by moving the histone octamer out of the way of an active polymerase and thereby decondensing the chromatin.

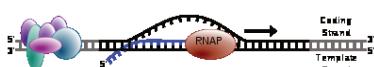
Transcription factors are modular in structure and contain the following domains:

DNA-binding domain (DBD), which attach to specific sequences of DNA enhancer or Promoter: Necessary component for all vectors: used to drive transcription of the vector's transgene promoter sequences) adjacent to regulated genes. DNA sequences that bind transcription factors are often referred to as **response elements**.

Trans-activating domain (TAD), which contain binding sites for other proteins such as [transcription coregulators. These binding sites are frequently referred to as **activation functions (AFs)**.^[30]

An optional **signal sensing domain (SSD)** (e.g., a ligand binding domain), which senses external signals and, in response, transmit these signals to the rest of the transcription complex, resulting in up- or down-regulation of gene expression. Also, the DBD and signal-sensing domains may reside on separate proteins that associate within the transcription complex to regulate gene expression.

Elongation of RNA



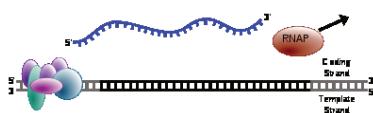
Simple diagram of transcription elongation (RNA in blue color)

After the first bond is synthesized, the RNA polymerase must clear the promoter. During this time there is a tendency to release

the RNA transcript and produce truncated transcripts. This is called abortive initiation and is common for both eukaryotes and prokaryotes. Abortive initiation continues to occur until the σ factor rearranges, resulting in the transcription elongation complex (which gives a 35 bp moving footprint). The σ factor is released before 80 nucleotides of mRNA are synthesized. Once the transcript reaches approximately 23 nucleotides, it no longer slips and elongation can occur. This, like most of the remainder of transcription, is an energy-dependent process, consuming adenosine triphosphate (ATP). One strand of the DNA, the template strand (or noncoding strand), is used as a template for RNA synthesis. As transcription proceeds, RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy. Although RNA polymerase traverses the template strand from $3' \rightarrow 5'$, the coding (non-template) strand and newly-formed RNA can also be used as reference points, so transcription can be described as occurring $5' \rightarrow 3'$. This produces an RNA molecule from $5' \rightarrow 3'$, an exact copy of the coding strand (except that thymines are replaced with uracils, and the nucleotides are composed of a ribose (5-carbon) sugar where DNA has deoxyribose (one less oxygen atom) in its sugar-phosphate backbone). Unlike DNA replication, mRNA transcription can involve multiple RNA polymerases on a single DNA template and multiple rounds of transcription (amplification of particular mRNA), so many mRNA molecules can be rapidly produced from a single copy of a gene. Elongation also involves a proofreading mechanism that can replace incorrectly incorporated bases. In eukaryotes, this

may correspond with short pauses during transcription that allow appropriate RNA editing factors to bind. These pauses may be intrinsic to the RNA polymerase or due to chromatin structure.^[31]

Transcription termination



Simple diagram of transcription termination. RNA is shown in blue color.

Bacteria use two different strategies for transcription termination. In Rho-independent transcription termination, RNA transcription stops when the newly synthesized RNA molecule

forms a G-C-rich hairpin loop followed by a run of Us. When the hairpin forms, the mechanical stress breaks the weak rU-dA bonds, now filling the DNA-RNA hybrid. This pulls the poly-U transcript out of the active site of the RNA polymerase, in effect, terminating transcription. In the "Rho-dependent" type of termination, a protein factor called "Rho" destabilizes the interaction between the template and the mRNA, thus releasing the newly synthesized mRNA from the elongation complex. Transcription termination in eukaryotes is less understood but involves cleavage of the new transcript followed by template-independent addition of As at its new 3' end, in a process called polyadenylation.

Rho-dependent termination

A Rho factor acts on an RNA substrate. Rho's key function is its helicase activity, for which energy is provided by an RNA-dependent ATP hydrolysis. The initial binding site for Rho is an extended (~70 nucleotides, sometimes 80-100 nucleotides) single-stranded region,

rich in cytosine and poor in guanine, called the rho utilization site or rut, in the RNA being synthesised, upstream of the actual terminator sequence. Several rho binding sequences have been discovered. No consensus is found among these, but the different sequences each seem specific, as small mutations in the sequence disrupts its function. Rho binds to RNA and then uses its ATPase activity to provide the energy to translocate along the RNA until it reaches the RNA-DNA helical region, where it unwinds the hybrid duplex structure. RNA polymerase pauses at the termination sequence, which is due to the fact that there is a specific site around 100nt away from the Rho binding site called the Rho-sensitive pause site. So, even though the RNA polymerase is about 40nt per second faster than Rho, it does not pose a problem for the Rho termination mechanism as the RNA polymerase allows Rho factor to catch up.

Rho-independent termination

Rho-independent termination (also known as Intrinsic termination) is a mechanism in both eukaryotes and prokaryotes that causes mRNA transcription to be stopped. In this mechanism, the mRNA contains a sequence that can base pair with itself to form a stem-loop structure 7-20 base pairs in length that is also rich in cytosine-guanine base pairs. These bases form three hydrogen bonds between each other and are therefore particularly strong. Following the stem-loop structure is a chain of uracil residues. The bonds between uracil and adenine are very weak. A protein bound to RNA polymerase (nusA) binds to the stem-loop structure tightly enough to cause the polymerase to temporarily stall. This pausing of the polymerase coincides with transcription of the poly-uracil sequence. The weak Adenine-Uracil bonds destabilize the RNA-DNA duplex, causing it to unwind and dissociate from the RNA polymerase. Stem-loop structures that are not followed by a poly-Uracil sequence cause the RNA polymerase to pause, but it will

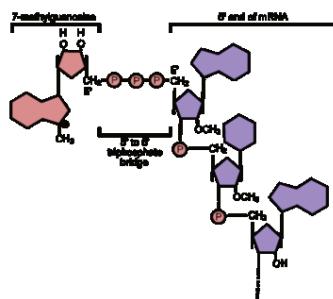
typically continue transcription after a brief time because the duplex is too stable to unwind far enough to cause termination. Rho-independent transcription termination is a frequent mechanism underlying the activity of cis-acting RNA regulatory elements, such as riboswitches^[32].

mRNA and its modification

The pre-mRNA molecule undergoes three main modifications. These modifications are 5' capping, 3' polyadenylation, and RNA splicing, which occur in the cell nucleus before the RNA is translated.

Capping of mRNA

The 5' cap looks like the 3' end of an RNA molecule (the 5' carbon of the cap ribose is bonded, and the 3' unbonded). This provides significant resistance to 5' exonucleases. Capping of the pre-mRNA involves the addition of 7-methylguanosine (m7G) to the 5' end. To achieve this, the terminal 5' phosphate requires



removal, which is done with the aid of a phosphatase enzyme. The enzyme guanosyl transferase then catalyses the reaction, which produces the diphosphate 5' end. The diphosphate 5' prime end then attacks the α phosphorus atom of a GTP molecule in order to add the guanine residue in a 5'5' triphosphate link. The enzyme

(guanine-N7)-methyltransferase (“cap MTase”) transfers a methyl group from S-adenosyl methionine to the guanine ring. This type of cap, with just the (m₇G) in position is called a cap 0 structure. The ribose of the adjacent nucleotide may also be methylated to give a cap 1. Methylation of nucleotides downstream of the RNA molecule produce cap 2, cap 3 structures and so on. In these cases the methyl groups are added to the 2' OH groups of the ribose sugar. The cap protects the 5' end of the primary RNA transcript from attack by ribonucleases that have specificity to the 3' phosphodiester bonds.^[33]

The starting point is the unaltered 5' end of an RNA molecule. This features a final nucleotide followed by three phosphate groups attached to the 5' carbon.

One of the terminal phosphate groups is removed (by RNA terminal phosphatase), leaving two terminal phosphates.

GTP is added to the terminal phosphates (by a guanylyl transferase), losing two phosphate groups (from the GTP) in the process. This results in the 5' to 5' triphosphate linkage.

The 7-nitrogen of guanine is methylated (by a methyl transferase).

Other methyltransferases are optionally used to carry out methylation of 5' proximal nucleotides.

The 5' cap has 4 main functions:

Regulation of nuclear export.

Prevention of degradation by exonucleases.

Promotion of translation (see ribosome and translation).

Promotion of 5' proximal intron excision.

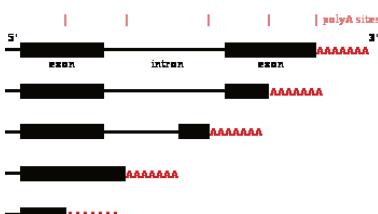
Nuclear export of RNA is regulated by the Cap binding complex (CBC), which binds exclusively to capped RNA. The CBC is then recognized by the nuclear pore complex and exported. Once in the cytoplasm after the pioneer round of translation, the CBC is replaced by the translation factors eIF-4E and eIF-4G. This complex is then recognized by other translation initiation machinery including the ribosome. Cap prevents 5' degradation in two ways. First, degradation of the mRNA by 5' exonucleases is prevented (as mentioned above) by functionally looking like a 3' end. Second, the

CBC complex and the eIF-4E/eIF-4G block the access of decapping enzymes to the cap. This increases the half-life of the mRNA, essential in eukaryotes as the export process takes significant time. Decapping of an mRNA is catalyzed by the decapping complex made up of at least Dcp1 and Dcp2, which must compete with eIF-4E to bind the cap. Thus the 5' cap is a marker of an actively translating mRNA and is used by cells to regulate mRNA half-lives in response to new stimuli. Undesirable mRNAs are sent to P-bodies for temporary storage or decapping, the details of which are still being resolved. The mechanism of 5' proximal intron excision promotion is not well understood, but the 5' cap appears to loop around and interact with the spliceosome in the splicing process, promoting intron excision.

Polyadenylation of mRNA

The pre-mRNA processing at the 3' end of the RNA molecule involves cleavage of its 3' end and then the addition of about 200 adenine residues to form a poly(A) tail. The cleavage and adenylation reactions occur if a polyadenylation signal sequence (5'- AAUAAA-3') is

located near the 3' end of the pre-mRNA molecule, which is followed by another sequence, which is usually (5'-CA-3'). The second signal is the site of cleavage. A **GU-rich sequence** is also usually present further downstream on the pre-mRNA molecule. After the synthesis of the sequence elements, two multisubunit proteins called cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CStF) are transferred from RNA Polymerase II to the RNA molecule. The two factors bind to the sequence elements.



Results of using different polyadenylation sites on the same gene

A protein complex forms that contains additional cleavage factors and the enzyme Polyadenylate Polymerase (PAP). This complex cleaves the RNA between the polyadenylation sequence and the GU-rich sequence at the cleavage site marked by the (5'-CA-3') sequences. Poly(A) polymerase then adds about 200 adenine units to the new 3' end of the RNA molecule using ATP as a precursor. As the poly(A) tail is synthesised, it binds multiple copies of poly(A) binding protein, which protects the 3' end from ribonuclease digestion.^[34]

Poly(A)-binding protein or “PABP”

Poly(A)-binding protein (or “PABP”) is a RNA-binding protein which binds to the poly(A) tail of mRNA. The poly(A) tail is located on the 3' end of mRNA. The nuclear isoforms selectively binds to around 50 nucleotides and stimulates the activity of Polyadenylate polymerase. The expression of mammalian Poly(A)-binding protein is regulated at the translational level by a feed-back mechanism: the mRNA encoding PABP contains in its 5' UTR an A-rich sequence which binds Poly(A)-binding protein. This leads to repression of translation. The cytosolic isoform of eukaryotes Poly(A) binding protein binds to the initiation factor eIF-4G via its C-terminal domain. EIF-4G is bound to eIF-4E, another initiation factor bound to the 5' cap on the 5' end of mRNA. This binding forms the characteristic loop structure of eukaryotic protein synthesis. Poly(A)-binding protein interacting proteins in the cytosol compete for the eIF-4G binding sites. Poly(A)-binding protein has also been shown to interact with a termination factor (eRF3).

Alternative polyadenylation

Many protein-coding genes have more than one polyadenylation site, so a gene can code for several mRNAs that differ in their 3' end. Since alternative polyadenylation changes the length of the 3' untranslated region, it can change which binding sites for microRNAs the 3' untranslated region contains. MicroRNAs tend to repress translation and promote degradation of the mRNAs they bind to, although there are examples of microRNAs that stabilise transcripts. Alternative polyadenylation can also shorten the coding

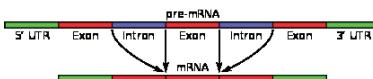
region, thus making the mRNA code for a different protein, but this is much less common than just shortening the 3' untranslated region. The choice of poly(A) site depends on the expression of the proteins that take part in polyadenylation. For example, the expression of CstF-64, a subunit of cleavage stimulatory factor (CstF), increases in macrophages in response to lipopolysaccharides (a group of bacterial compounds that trigger an immune response). This results in the selection of weak poly(A) sites and thus shorter transcripts. This removes regulatory elements in the 3' untranslated regions of mRNAs for defense-related products like lysozyme and TNF- α . These mRNAs then have longer half-lives and produce more of these proteins. RNA-binding proteins other than those in the polyadenylation machinery can also affect whether a polyadenylation site is used, as can DNA methylation near the polyadenylation signal^[35].

Splicing of RNA

Splicing is the process by which pre-mRNA is modified to remove certain stretches of non-coding sequences called introns; the stretches that remain include protein-coding sequences and are called exons.

Sometimes pre-mRNA messages may be spliced in several different ways, allowing a single gene to encode multiple proteins. This process is called alternative splicing. Splicing is usually performed by an RNA-protein complex called the spliceosome, but some RNA molecules are also capable of catalyzing their own splicing^[36]

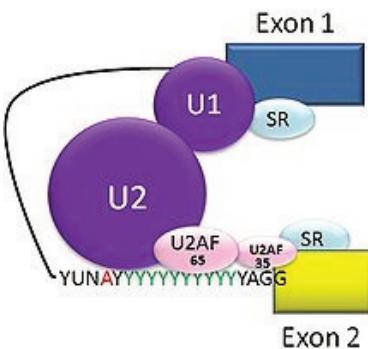
Introns



Simple illustration of a pre-mRNA, with introns (top). After the introns have been removed via splicing, the mature mRNA sequence is ready for translation (bottom).

One plausible hypothesis for the observed distribution of introns is that ancient predecessors of modern-day eukaryotes contained large numbers of introns, and that selective pressure to control genome size in fast-growing species may have led to the elimination of many ancient introns. Complicating this issue is that finding that many introns are themselves mobile genetic elements, and can be inserted into and deleted from genes. Shortly after the discovery of introns, investigators offered competing theories that offer alternative scenarios for the origin and early evolution of spliceosomal introns. Other classes of introns such as self-splicing and tRNA introns are not subject to much debate, but see for the former. These are popularly referred to as the Introns-Early (IE) and the Introns-Late (IL) views. The IE model, championed by Walter Gilbert, proposes that introns are extremely old and numerously present in the earliest theoretical ancestors of prokaryotes and eukaryotes, the progenotes. In this model, introns were subsequently lost from prokaryotic organisms, allowing them to attain growth efficiency. A central prediction of this theory is that the early introns were mediators that facilitated the recombination of exons that represented the protein domains. This model cannot account for some observed positional variation of introns shared among related genes.

The IL model proposes that introns were recently inserted into originally intron-less contiguous genes after the divergence of eukaryotes and prokaryotes. In this model, introns probably originated from transposable elements. This model is based on the observation that the spliceosomal introns are restricted to



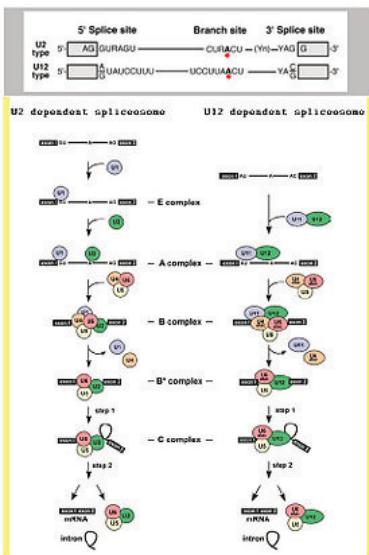
Spliceosome A complex defines the 5' and 3' ends of the intron before removal

eukaryotes alone. However, there is considerable debate over the presence of introns in the early prokaryote-eukaryote ancestors and the subsequent intron loss-gain during eukaryotic evolution. The evolution of introns and of the intron-exon structure may be largely independent of the evolution of coding-sequences.

Introns were first discovered in protein-coding genes of adenovirus , but are now known to occur within a wide variety of genes throughout all of the biological kingdoms. The frequency of introns within different genomes can vary widely across the spectrum of biological organisms. For example, introns are extremely common within the nuclear genome of higher vertebrates (e.g. humans and mice), where protein-coding genes almost always contain multiple introns, while introns are rare within the nuclear genes of some eukaryotic microorganisms, for example baker's yeast (*Saccharomyces cerevisiae*). In contrast, the mitochondrial genomes of vertebrates are entirely devoid of introns, while those of eukaryotic microorganisms may contain many introns.

Spliceosome and its assembly

Each spliceosome is composed of five small nuclear RNA proteins, called snRNPs, (pronounced “snurps”) and a range of non-snRNP associated protein factors. The snRNPs that make up the nuclear spliceosome are named U1, U2, U4, U5, and U6, and participate in several RNA-RNA and RNA-protein interactions. The RNA component of the snRNP is rich in uridine (the nucleoside analog of the uracil nucleotide). The canonical assembly of the spliceosome occurs anew on each hnRNA. The hnRNA contains specific sequence elements that are recognized and utilized during spliceosome assembly. These include the 5' end splice, the branch point sequence, the polypyrimidine tract, and the 3' end splice site. The spliceosome catalyzes the removal of introns, and the ligation of the flanking exons. Introns typically have a GU nucleotide sequence at the 5' end splice site, and an AG at the 3' end splice site. The 3' splice site can be further defined by a variable length of polypyrimidines, called the polypyrimidine tract (PPT), which serves the dual function of recruiting factors to the 3' splice site and possibly recruiting factors to the branch point sequence (BPS). The BPS contains the conserved Adenosine required for the first step of splicing. A group of less abundant snRNPs, U11, U12, U4atac, and U6atac, together with U5, are subunits of the so-called minor spliceosome that splices a rare class of pre-mRNA introns,



U5 is believed to be the only common component between major and minor spliceosomes.
Reference: Pelli G, International J Biol Chem. 2005 Aug;36(8):712-24.

A comparison between major and minor splicing mechanisms^[37]

denoted U12-type. These snRNPs form the U12 spliceosome are located in the cytosol. New evidence derived from the first crystal structure of a group II intron suggests that the spliceosome is actually a ribozyme, and that it uses a two-metal ion mechanism for catalysis. The model for formation of the spliceosome active site involves an ordered, stepwise assembly of discrete snRNP particles on the hnRNA substrate. The first recognition of hnRNAs involves U1 snRNP binding to the 5' end splice site of the hnRNA and other non-snRNP associated factors to form the commitment complex, or early (E) complex in mammals.^{[38][39]} The commitment complex is an ATP-independent complex that commits the hnRNA to the splicing pathway.^[40] U2 snRNP is recruited to the branch region through interactions with the E complex component U2AF (U2 snRNP auxiliary factor) and possibly U1 snRNP. In an ATP-dependent reaction, U2 snRNP becomes tightly associated with the branch point sequence (BPS) to form complex A. A duplex formed between U2 snRNP and the hnRNA branch region bulges out the branch adenosine specifying it as the nucleophile for the first transesterification.^[41] The presence of a pseudouridine residue in U2 snRNA, nearly opposite of the branch site, results in an altered conformation of the RNA-RNA duplex upon the U2 snRNP binding. Specifically, the altered structure of the duplex induced by the pseudouridine places the 2' OH of the bulged adenosine in a favorable position for the first step of splicing.^[42] The U4/U5/U6 tri-snRNP is recruited to the assembling spliceosome to form complex B, and following several rearrangements, complex C (the spliceosome) is activated for catalysis.^{[43][44]} It is unclear how the triple snRNP is recruited to complex A, but this process may be mediated through protein-protein interactions and/or base pairing interactions between U2 snRNA and U6 snRNA. The U5 snRNP interacts with sequences at the 5' and 3' splice sites via the invariant loop of U5 snRNA^[45] and U5 protein components interact with the 3' splice site region.^[46] Upon recruitment of the triple snRNP, several RNA-RNA rearrangements precede the first catalytic step and further rearrangements occur in the catalytically active

spliceosome. Several of the RNA-RNA interactions are mutually exclusive; however, it is not known what triggers these interactions, nor the order of these rearrangements. The first rearrangement is probably the displacement of U1 snRNP from the 5' splice site and formation of a U6 snRNA interaction. It is known that U1 snRNP is only weakly associated with fully formed spliceosomes^[47], and U1 snRNP is inhibitory to the formation of a U6-5' splice site interaction on a model of substrate oligonucleotide containing a short 5' exon and 5' splice site.^[48] Binding of U2 snRNP to the branch point sequence (BPS) is one example of an RNA-RNA interaction displacing a protein-RNA interaction. Upon recruitment of U2 snRNP, the branch binding protein SF1 in the commitment complex is displaced since the binding site of U2 snRNA and SF1 are mutually exclusive events. Within the U2 snRNA, there are other mutually exclusive rearrangements that occur between competing conformations. For example, in the active form, stem loop IIa is favored; in the inactive form a mutually exclusive interaction between the loop and a downstream sequence predominates.^[44] It is unclear how U4 is displaced from U6 snRNAm, although RNA has been implicated in spliceosome assembly, and may function to unwind U4/U6 and promote the formation of a U2/U6 snRNA interaction. The interactions of U4/U6 stem loops I and II dissociate and the freed stem loop II region of U6 folds on itself to form an intramolecular stem loop and U4 is no longer required in further spliceosome assembly. The freed stem loop I region of U6 base pairs with U2 snRNA forming the U2/U6 helix I. However, the helix I structure is mutually exclusive with the 3' half of an internal 5' stem loop region of U2 snRNA. The RNA components of snRNPs interact with the intron and may be involved in catalysis. Two types of spliceosomes have been identified (the major and minor) which contain different snRNPs. Major The major spliceosome splices introns containing GU at the 5' splice site and AG at the 3' splice site. It is composed of the U1, U2, U4, U5, and U6 snRNPs and is active in the nucleus. In addition, a number of proteins including U2AF and SF1 are required for the assembly of the spliceosome. E Complex-U1

binds to the GU sequence at the 5' splice site, along with accessory proteins/enzymes ASF/SF2, U2AF (binds at the Py-AG site), SF1/BBP (BBP=Branch Binding Protein); A Complex-U2 binds to the branch site and ATP is hydrolyzed; B1 Complex-U5/U4/U6 trimer binds, and the U5 binds exons at the 5' site, with U6 binding to U2; B2 Complex-U1 is released, U5 shifts from exon to intron and the U6 binds at the 5' splice site;

C1 Complex-U4 is released, U6/U2 catalyzes transesterification, that make 5'end of introns ligate to the A on intron and from a lariat, U5 binds exon at 3' splice site, and the 5' site is cleaved, resulting in the formation of the lariat;

C2 Complex-U2/U5/U6 remain bound to the lariat, and the 3' site is cleaved and exons are ligated using ATP hydrolysis. The spliced RNA is released and the lariat debranches.

This type of splicing is termed canonical splicing or termed the lariat pathway, which accounts for more than 99% of splicing. By contrast, when the intronic flanking sequences do not follow the GU-AG rule, noncanonical splicing is said to occur (see “minor spliceosome” below).

Minor spliceosome

The minor spliceosome is a ribonucleoprotein complex that catalyses the removal (splicing) of an atypical class of spliceosomal introns (U12-type) from eukaryotic messenger RNAs in plant, insects, vertebrates and some fungi (*Rhizopus oryzae*). This process is called noncanonical splicing, as opposed to U2-dependent canonical splicing. U12-type introns represent less than 1% of all introns in human cells. However they are found in genes performing essential cellular functions. The minor spliceosome is very similar to the major spliceosome, however it splices out rare introns with different splice site sequences. While the minor and major spliceosomes contain the same U5 snRNP, the minor spliceosome has different, but functionally analogous snRNPs for U1, U2, U4, and U6, which are respectively called U11, U12, U4atac, and U6atac. Like the major spliceosome, it is only found in the nucleus^[49]. Trans-

splicing Trans-splicing is a form of splicing that joins two exons that are not within the same RNA transcript

Self splicing

Self splicing occurs for rare introns that form a ribozyme, performing the functions of the spliceosome by RNA alone. There are three kinds of self-splicing introns, Group I, Group II and Group III. Group I and II introns perform splicing similar to the spliceosome without requiring any protein. This similarity suggests that Group I and II introns may be evolutionarily related to the spliceosome. Self-splicing may also be very ancient, and may have existed in an RNA world present before protein. Two transesterifications characterize the mechanism in which group I introns are spliced: 3'OH of a free guanine nucleoside (or one located in the intron) or a nucleotide cofactor (GMP, GDP, GTP) attacks phosphate at the 5' splice site. 3'OH of the 5' exon becomes a nucleophile and the second transesterification results in the joining of the two exons. The mechanism in which group II introns are spliced (two transesterification reaction like group I introns) is as follows: The 2'OH of a specific adenosine in the intron attacks the 5' splice site, thereby forming the lariat. The 3'OH of the 5' exon triggers the second transesterification at the 3' splice site thereby joining the exons together.

Group I

Group I introns are distributed in bacteria, lower eukaryotes and higher plants. However, their occurrence in bacteria seems to be more sporadic than in lower eukaryotes, and they have become prevalent in higher plants. The genes that group I introns interrupt differ significantly: They interrupt rRNA, mRNA and tRNA genes in bacterial genomes, as well as in mitochondrial and chloroplast genomes of lower eukaryotes, but only invade rRNA genes in the nuclear genome of higher plants, these introns

seem to be restricted to a few tRNA and mRNA genes of the chloroplasts and mitochondria. Both intron-early and intron-late theories have found evidences in explaining the origin of group I introns. Some group I introns encode homing endonuclease (HEG), which catalyzes intron mobility. It is proposed that HEGs move the intron from one location to another, from one organism to another and thus account for the wide spreading of the selfish group I introns. No biological role has been identified for group I introns thus far except for splicing of themselves from the precursor to prevent the death of the host that they live by. A small number of group I introns are also found to encode a class of proteins called maturases that facilitate the intron splicing.

Splicing of group I introns is processed by two sequential ester-transfer reactions. The exogenous guanosine or guanosine nucleotide (exoG) first docks onto the active G-binding site located in P7, and its 3'-OH is aligned to attack the phosphodiester bond at the 5' splice site located in P1, resulting in a free 3'-OH group at the upstream exon and the exoG being attached to the 5' end of the intron. Then the terminal G (omega G) of the intron swaps the exoG and occupies the G-binding site to organize the second ester-transfer reaction, the 3'-OH group of the upstream exon in P1 is aligned to attacks the 3' splice site in P10, leading to the ligation of the adjacent upstream and downstream exons and free of the catalytic intron. Two-metal-ion mechanism seen in protein polymerases and phosphatases was proposed to be used by group I and group II introns to process the phosphoryl transfer reactions, which was unambiguously proven by a recently resolved high-resolution structure of the *Azoarcus* group I intron.

Group II catalytic intron

Group II catalytic introns are found in rRNA, tRNA and mRNA of organelles in fungi, plants and protists, and also in mRNA in bacteria. They are large self-splicing ribozymes and have 6 structural domains (usually designated dI to dVI). This model and alignment represents only domains V and VI. A subset of group II introns also encode essential splicing proteins in intronic ORFs. The length

of these introns can therefore be up to 3kb. Splicing occurs in almost identical fashion to nuclear pre-mRNA splicing with two transesterification steps. The 2' hydroxyl of a bulged adenosine in domain VI attacks the 5' splice site, followed by nucleophilic attack on the 3' splice site by the 3' OH of the upstream exon. Protein machinery is required for splicing *in vivo*, and long range intron-intron and intron-exon interactions are important for splice site positioning. Group II introns are further sub-classified into groups IIA and IIB, which differ in splice site consensus, and the distance of the bulged adenosine in domain VI (the prospective branch point forming the lariat) from the 3' splice site.

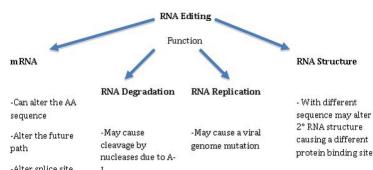
Group III introns

Montandon,P. and Stutz,E. (1984) and Hallick,R.B. et al. (1988 and 1989) reported examples of a novel type of introns in Euglena chloroplast. In 1989, David A.Christopher and Richard B.Hallick proposed the title, Group III introns to identify this new class with the following characteristics: Group III introns are much shorter than other self-splicing intron classes, ranging from 95 to 110 nucleotides amongst those known to Christopher and Hallick, and identified in chloroplasts. On the other hand, Christopher and Hallick stated: "By contrast, the smallest Euglena chloroplast group II intron ... is 277 nucleotides." Their conserved sequences proximal to the splicing sites have similarities to those of group II introns, but have fewer conserved positions. They do not map into the conserved secondary structure of group II introns. (Indeed Christopher and Hallick were unable to identify any conserved secondary structure elements among group III introns.) They are usually associated with genes involved in translation and transcription. They are very A+T rich. In 1994, discovery of a group III intron with a length of one order of magnitude longer indicated that length alone is not the determinant of splicing in Group III introns (Copertino DW., Hall ET.) Splicing of group III introns occurs through lariat and circular RNA formation. Similarities between group III and nuclear introns include conserved 5' boundary

sequences, lariat formation, lack of internal structure, and ability to use alternate splice boundaries.

RNA-editing

The RNA-editing system seen in the animal may have evolved from mononucleotide deaminases, which have led to larger gene families that include the apobec-1 and adar genes. These genes share close identity with the bacterial deaminases involved in nucleotide metabolism. The adenosine deaminase of *E. coli* cannot deaminate a nucleoside in the RNA; the enzyme's reaction pocket is too small to for the RNA strand to bind to. However, this active site is widened by amino acid changes in the corresponding human analog genes, APOBEC-1 and ADAR, allowing deamination. The insertional editing seen in the trypanosome mitochondria has no relation with the nucleoside conversion process. The enzymes involved have been shown in other studies to be recruited and adapted from different sources. But, the specificity of nucleotide insertion via the interaction between the gRNA and mRNA are similar to the tRNA editing processes in the animal and *Acanthamoeba* mitochondria. Furthermore, the eukaryotic ribose methylation of rRNAs by guide RNA molecules may provide another link between RNA editing and modification. As a consequence, the numerous studies suggest that RNA editing may have evolved in specific lineages of speciation, due to the subtle differences in their mechanism. The data does not support the existence of RNA editing in the RNA world, since its mechanism is not linked to any hypothesized process that may have existed at that time. Therefore,



Summary of the Various Functions of RNA Editing

RNA editing appears to have evolved at a later time to compensate for the changes in gene sequences and to increase variation.

Editing by insertion or deletion

RNA editing through the addition and deletion of uracil has been found in kinetoplasts from the mitochondria of *Trypanosoma brucei*. Editing of the RNA starts with the base-pairing of the unedited primary transcript with a guide RNA (gRNA), which contains complementary sequences to the regions around the insertion/deletion points. The newly formed double-stranded region is then enveloped by an editosome, a large multi-protein complex that catalyzes the editing. The editosome opens the transcript at the first mismatched nucleotide and starts inserting uridines. The inserted uridines will base-pair with the guide RNA, and insertion will continue as long as A or G is present in the guide RNA and will stop when a C or U is encountered. The inserted nucleotides cause a frameshift and result in a translated protein that differs from its gene.

The Editosome Complex

The mechanism of the editosome involves an endonucleolytic cut at the mismatch point between the guide RNA and the unedited transcript. The next step is catalyzed by one of the enzymes in the complex, a terminal U-transferase, which adds Us from UTP at the 3' end of the mRNA. The opened ends are held in place by other proteins in the complex. Another enzyme, a U-specific exoribonuclease, removes the unpaired Us. After editing has made mRNA complementary to gRNA, an RNA ligase rejoins the ends of the edited mRNA transcript. As a consequence, the editosome can edit only in a 3' to 5' direction along the primary RNA transcript. The complex can act on only a single guide RNA at a time. Therefore, a RNA transcript requiring extensive editing will need more than one guide RNA and editosome complex.

Editing by deamination

C-U editing

The editing involves cytidine deaminase that deaminates a cytidine base into a uridine base. An example of C-to-U editing is with the apolipoprotein B gene in humans. Apo B100 is expressed in the liver and apo B48 is expressed in the intestines. The B100 form has a CAA sequence that is edited to UAA, a stop codon, in the intestines. It is unedited in the liver.

A-I editing

A-to-I editing occurs in regions of double-stranded RNA (dsRNA). Adenosine deaminases acting on RNA (ADARs) are the RNA-editing enzymes involved in the hydrolytic deamination of Adenosine to Inosine (A-to-I editing). A-to-I editing can be specific (a single adenosine is edited within the stretch of dsRNA) or promiscuous (up to 50% of the adenosines are edited). Specific editing occurs within short duplexes (e.g., those formed in an mRNA where intronic sequence base pairs with a complementary exonic sequence), while promiscuous editing occurs within longer regions of duplex (e.g., pre- or pri-miRNAs, duplexes arising from transgene or viral expression, duplexes arising from paired repetitive elements). There are many effects of A-to-I editing, arising from the fact that I behaves as if it is G both in translation and when forming secondary structures. These effects include alteration of coding capacity, altered miRNA or siRNA target populations, heterochromatin formation, nuclear sequestration, cytoplasmic sequestration, endonucleolytic cleavage by Tudor-SN, inhibition of miRNA and siRNA processing ,and altered splicing^[50].

References and online resources

Ross Hardison clears up the contrary usages of the terms “sense strand” and “coding strand”. See his short free chapter on

Transcription and mRNA structure from his textbook Working With Molecular Genetics.

References

1. ↑ Littlefield, O., Korkhin, Y., and Sigler, P.B. (1999). “The structural basis for the oriented assembly of a TBP/TFB/promoter complex”. *PNAS* **96** (24): 13668–13673. doi:10.1073/pnas.96.24.13668. PMID 10570130.
2. ↑ Hausner, W; Thomm, M (2001). “Events during Initiation of Archaeal Transcription: Open Complex Formation and DNA-Protein Interactions”. *Journal of Bacteriology* **183** (10): 3025–3031. doi:10.1128/JB.183.10.3025-3031.2001. PMID 11325929.
3. ↑ Qureshi, SA; Bell, SD; Jackson, SP (1997). “Factor requirements for transcription in the archaeon Sulfolobus shibatae”. *EMBO Journal* **16** (10): 2927–2936. doi:10.1093/emboj/16.10.2927. PMID 9184236.
4. ↑ TATA-binding protein
5. ↑ Jerard Hurwitz (December 2005). “The Discovery of RNA Polymerase”. *Journal of Biological Chemistry* **280** (52): 42477–85. doi:10.1074/jbc.X500006200. PMID 16230341.
6. ↑ Grummt I. (1999). “Regulation of mammalian ribosomal gene transcription by RNA polymerase I.”. *Prog Nucleic Acid Res Mol Biol.* **62**: 109–54. doi:10.1016/S0079-6603(08)60506-1. PMID 9932453.
7. ↑ Lee Y; Kim M; Han J; Yeom KH; Lee S; Baek SH; Kim VN. (October 2004). “MicroRNA genes are transcribed by RNA polymerase II”. *EMBO J.* **23** (20): 4051–60. doi:10.1038/sj.emboj.7600385. PMID 15372072.
8. ↑ Sawadogo M, Sentenac A (1990). “RNA polymerase B (II) and general transcription factors.”. *Annu Rev Biochem.* **59**: 711–54.

- doi:10.1146/annurev.bi.59.070190.003431. PMID 2197989.
- 9. ↑ Myer VE, Young RA (October 1998). “RNA polymerase II holoenzymes and subcomplexes”. *J. Biol. Chem.* **273** (43): 27757–60. doi:10.1074/jbc.273.43.27757. PMID 9774381.
<http://www.jbc.org/cgi/reprint/273/43/27757.pdf>.
 - 10. ↑ Jump up to: **a b c d e f g h** Acker J, de Graaff M, Cheynel I, Khazak V, Kedinger C, Vigneron M (Jul 1997). “Interactions between the human RNA polymerase II subunits”. *J Biol Chem.* **272** (27): 16815–21. doi:10.1074/jbc.272.27.16815. PMID 9201987.
 - 11. ↑ Brickey WJ, Greenleaf AL (June 1995). “Functional studies of the carboxy-terminal repeat domain of Drosophila RNA polymerase II in vivo”. *Genetics* **140** (2): 599–613. PMID 7498740.
 - 12. ↑ “Entrez Gene: POLR2A polymerase (RNA) II (DNA directed) polypeptide A, 220kDa”. <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd>ShowDetailView&TermToSearch=5430>
 - 13. ↑ “Entrez Gene: POLR2B polymerase (RNA) II (DNA directed) polypeptide B, 140kDa”. <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd>ShowDetailView&TermToSearch=5431>.
 - 14. ↑ Khazak V, Estojak J, Cho H, Majors J, Sonoda G, Testa JR, Golemis EA (May 1998). “Analysis of the interaction of the novel RNA polymerase II (pol II) subunit hsRPB4 with its partner hsRPB7 and with pol II”. *Mol Cell Biol.* **18** (4): 1935–45. PMID 9528765.
 - 15. ↑ “Entrez Gene: POLR2E polymerase (RNA) II (DNA directed) polypeptide E, 25kDa”. <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd>ShowDetailView&TermToSearch=5434>.
 - 16. ↑ “Entrez Gene: POLR2F polymerase (RNA) II (DNA directed) polypeptide F”. <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd>ShowDetailView&TermToSearch=5435>.
 - 17. ↑ “Entrez Gene: POLR2G polymerase (RNA) II (DNA directed) polypeptide G”. <http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd>ShowDetailView&TermToSearch=5436>.

18. ↑ “POLR2J3 polymerase (RNA) II (DNA directed) polypeptide J3”. http://www.ncbi.nlm.nih.gov/gene/548644?ordinalpos=1&itool=EntrezSystem2.PEntrez.Gene.Gene_ResultsPanel.Gene_RVDocSum.
19. ↑ Brickey WJ, Greenleaf AL (June 1995). “Functional studies of the carboxy-terminal repeat domain of Drosophila RNA polymerase II in vivo”. *Genetics*. **140** (2): 599–613. PMID 7498740.
20. ↑ Meinhart A, Cramer P (Jul 2004). “Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors” (abstract). *Nature*. **430** (6996): 223–6. doi:10.1038/nature02679. PMID 15241417. <http://www.nature.com/nature/journal/v430/n6996/abs/nature02679.htm>.
21. ↑ Jump up to: **a b** Corden JL (1990). “Tails of RNA polymerase II”. *Trends Biol Sci*. **15**: 383–7. doi:10.1016/0968-0004(90)90236-5.
22. ↑ Jump up to: **a b** Phatnani HP, Greenleaf AL (Nov 2006). “Phosphorylation and functions of the RNA polymerase II CTD”. *Genes Dev*. **20** (1): 2922–36. doi:10.1101/gad.1477006. PMID 17079683. <http://genesdev.cshlp.org/content/20/21/2922.long>.
23. ↑ http://en.wikipedia.org/wiki/RNA_polymerase_II_holoenzyme
24. ↑ http://en.wikipedia.org/wiki/RNA_polymerase_II_holoenzyme
25. ↑ Willis IM. (February 1993). “RNA polymerase III. Genes, factors and transcriptional specificity”. *Eur J Biochem*. **212** (1): 1–11. doi:10.1111/j.1432-1033.1993.tb17626.x. PMID 8444147.
26. ↑ Herr AJ, Jensen MB, Dalmay T, Baulcombe DC (2005). “RNA polymerase IV directs silencing of endogenous DNA”. *Science* **308** (5718): 118–20. doi:10.1126/science.1106910. PMID 15692015.
27. ↑ Wierzbicki AT, Ream TS, Haag JR, Pikaard CS (May 2009). “RNA polymerase V transcription guides ARGONAUTE4 to chromatin”. *Nat. Genet*. **41** (5): 630–4. doi:10.1038/ng.365. PMID 19377477.
28. ↑ Makeyev EV, Bamford DH (December 2002). “Cellular RNA-

- dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes". *Mol. Cell* **10** (6): 1417–27. doi:10.1016/S1097-2765(02)00780-3. PMID 12504016.
<http://linkinghub.elsevier.com/retrieve/pii/S1097276502007803>.
29. ↑ Mohamed Ouhammouch, Robert E. Dewhurst, Winfried Hausner, Michael Thomm, and E. Peter Geiduschek (2003). "Activation of archaeal transcription by recruitment of the TATA-binding protein". *Proceedings of the National Academy of Sciences of the United States of America* **100** (9): 5097–5102. doi:10.1073/pnas.0837150100. PMID 12692306.
30. ↑ Wärnmark A, Treuter E, Wright AP, Gustafsson J-Å (2003). "Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation". *Mol. Endocrinol.* **17** (10): 1901–9. doi:10.1210/me.2002-0384. PMID 12893880.
31. ↑ Transcription (genetics)
32. ↑ http://en.wikipedia.org/w/index.php?title=Intrinsic_termination&oldid=428278551
33. ↑ 5'cap
34. ↑ Hames & Hooper 2006, p. 225
35. ↑ <http://en.wikipedia.org/w/index.php?title=Polyadenylation&oldid=422375377>
36. ↑ Messenger_RNA
37. ↑ Minor spliceosome
38. ↑ Jamison SF, Crow A, and Garcia-Blanco MA (October 1, 1992). "The Spliceosome Assembly Pathway in Mammalian Extracts". *Molecular and Cell Biology* **12** (10): 4279–87. PMID 1383687.
39. ↑ Seraphin B. and Rosbash M. (1989). "Identification of functional U1 snRNA pre-messenger RNA complexes committed to spliceosome assembly and splicing". *Cell* **59** (2): 349–58. doi:10.1016/0092-8674(89)90296-1. PMID 2529976.
40. ↑ Legrain P, Seraphin B, Rosbash M (September 1, 1988). "Early commitment of yeast pre-mRNA to the spliceosome pathway". *Mol. Cell. Biol.* **8** (9): 3755–60. PMID 3065622. PMC 365433.
<http://mcb.asm.org/cgi/reprint/8/9/3755>.

41. ↑ Query, C. C., M. J. Moore, and P. Sharp (1994). “Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model”. *Genes Devel.* **8** (5): 587–97. doi:10.1101/gad.8.5.587. PMID 7926752. <http://www.genesdev.org/cgi/pmidlookup?view=long&pmid=7926752>.
42. ↑ Newby M. I. and Greenbaum, N. L. (2002). “Sculpting of the spliceosomal branch site recognition motif by a conserved pseudouridine”. *Nature Structural Biology* **9** (12): 958–65. doi:10.1038/nsb873. PMID 12426583.
43. ↑ Burge, C.B., et al. (1999). “Splicing precursors to mRNAs by the spliceosomes”. in Gesteland, R.F., Cech, T.R., Atkins, J.F.. *The RNA World*. Cold Spring Harbor Lab. Press. pp. 525–60. ISBN 0879693800.
44. ↑ ^{Jump up to: **a b**} Staley JP, Guthrie C (1998). “Mechanical devices of the spliceosome: motors, clocks, springs, and things”. *Cell* **92** (3): 315–26. doi:10.1016/S0092-8674(00)80925-3. PMID 9476892.
45. ↑ Newman AJ, Teigelkamp S and Beggs JD (1995). “snRNA interactions at 5' and 3' splice sites monitored by photoactivated crosslinking in yeast spliceosomes”. *RNA* **1** (9): 968–80. PMID 8548661. PMC 1369345. <http://www.rnajournal.org/cgi/reprint/1/9/968>.
46. ↑ Chiara MD, Palandjian L, Feld Kramer R, Reed R (1997). “Evidence that U5 snRNP recognizes the 3' splice site for catalytic step II in mammals”. *EMBO J.* **16** (15): 4746–59. doi:10.1093/emboj/16.15.4746. PMID 9303319. PMC 1170101. <http://www.nature.com/emboj/journal/v16/n15/abs/7590453a.html>.
47. ↑ Moore, M. J. and Sharp, P. A. (1993). “Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing”. *Nature* **365** (6444): 364–8. doi:10.1038/365364a0. PMID 8397340.
48. ↑ Konforti BB, Koziolkiewicz MJ, Konarska MM (1993). “Disruption of base pairing between the 5' splice site and the 5' end of U1 snRNA is required for spliceosome assembly”. *Cell* **75**

- (5): 863–73. doi:10.1016/0092-8674(93)90531-T. PMID 8252623.
49. ↑ Will CL, Lührmann R (August 2005). “Splicing of a rare class of introns by the U12-dependent spliceosome”. *Biol. Chem.* 386 (8): 713–24
50. ↑ http://en.wikipedia.org/w/index.php?title=RNA_editing&oldid=424324287

I6.

After the structure of DNA was discovered by James Watson and Francis Crick, who used the experimental evidence of Maurice Wilkins and Rosalind Franklin (among others), serious efforts to understand the nature of the encoding of proteins began. George Gamow, in 1954,^[1] postulated that a three-letter code must be employed to encode the 20 standard amino acids used by living cells to encode protein. Three is the smallest integer n such that 4^n is at least 20.

The fact that codons consist of three DNA bases was first demonstrated in the Crick, Brenner et al. experiment.^[2] **The first elucidation of a codon was done by Marshall Nirenberg and Heinrich J. Matthaei in 1961 at the National Institutes of Health.**^[3] They used a cell-free system to translate a poly-uracil RNA sequence (i.e., UUUUU...) and discovered that the polypeptide that they had synthesized consisted of only the amino acid phenylalanine. They thereby deduced that the codon UUU specified the amino acid phenylalanine. This was followed by experiments in the laboratory of **Severo Ochoa** demonstrating that the **poly-adenine RNA sequence** (AAAAA...) coded for the polypeptide, poly-lysine. and the **poly-cytosine RNA sequence** (CCCCC...) coded for the polypeptide, poly-proline. Therefore the codon AAA specified the amino acid lysine, and the codon CCC specified the amino acid proline. Using different copolymers most of the remaining codons were then determined. Extending this work, Nirenberg and Philip Leder revealed the triplet nature of the genetic code and allowed the codons of the standard genetic code to be deciphered. In these experiments various combinations of mRNA were passed through a filter which contained ribosomes, the components of cells that translate RNA into protein. Unique triplets promoted the binding of specific tRNAs to the ribosome. **Leder and Nirenberg were able**

to determine the sequences of 54 out of 64 codons in their experiments.

Subsequent work by **Har Gobind Khorana** identified the rest of the genetic code. Shortly thereafter, **Robert W. Holley determined the structure of transfer RNA (tRNA)**, the adapter molecule that facilitates the process of translating RNA into protein. This work was based upon earlier studies by Severo Ochoa, who received the Nobel prize in 1959 for his work on the enzymology of RNA synthesis. **In 1968, Khorana, Holley and Nirenberg received the Nobel Prize in Physiology or Medicine for their work.**^[4]

Contents

- 1 Origin of genetic code
- 2 Summary of Khorana's research
 - 2.1 The table of Genetic Code
- 3 Degeneracy of the genetic code
- 4 Initiation and Termination codon
 - 4.1 Initiation or Start Codon
 - 4.2 Termination or Stop codon
- 5 Facts to be remembered
- 6 References

Origin of genetic code

There are many theories behind the origin of genetic codes. The genetic code used by all known forms of life is nearly universal. However, there are a huge number of possible genetic codes. If amino acids are randomly associated with triplet codons, there will be 1.5×10^{84} possible genetic codes. Phylogenetic analysis of

transfer RNA suggests that tRNA molecules evolved before the present set of aminoacyl-tRNA synthetases.

Theoretically the genetic code could be completely random (a “frozen accident”), completely non-random (optimal) or a combination of random and nonrandom. There are sufficient data to refute the first possibility. For a start, a quick view on the table of the genetic code already shows a clustering of amino acid assignments. Furthermore, amino acids that share the same biosynthetic pathway tend to have the same first base in their codons, and amino acids with similar physical properties tend to have similar codons.

There are four themes running through the many theories that seek to explain the evolution of the genetic code (and hence the origin of these patterns):

1. Chemical principles govern specific RNA interaction with amino acids. Aptamer experiments showed that some amino acids have a selective chemical affinity for the base triplets that code for them. Recent experiments show that of the 8 amino acids tested, 6 show some RNA triplet-amino acid association. This has been called the stereochemical code. The stereochemical code could have created an ancient core of assignments. The current complex translation mechanism involving tRNA and associated enzymes may be a later development, and that originally, protein sequences were directly templated on base sequences.
2. Biosynthetic expansion. The standard modern genetic code grew from a simpler earlier code through a process of “biosynthetic expansion”. Here the idea is that primordial life “discovered” new amino acids (e.g., as by-products of metabolism) and later back-incorporated some of these into the machinery of genetic coding. Although much circumstantial evidence has been found to suggest that fewer different amino acids were used in the past than today, precise and detailed hypotheses about exactly which amino acids

entered the code in exactly what order have proved far more controversial.

3. Natural selection has led to codon assignments of the genetic code that minimize the effects of mutations. A recent hypothesis suggests that the triplet code was derived from codes that used longer than triplet codons. Longer than triplet decoding has higher degree of codon redundancy and is more error resistant than the triplet decoding. This feature could allow accurate decoding in the absence of highly complex translational machinery such as the ribosome.
4. Information channels: Information-theoretic approaches see the genetic code as an error-prone information channel. The inherent noise (i.e. errors) in the channel poses the organism with a fundamental question: how to construct a genetic code that can withstand the impact of noise while accurately and efficiently translating information? These “rate-distortion” models suggest that the genetic code originated as a result of the interplay of the three conflicting evolutionary forces: the needs for diverse amino-acids, for error-tolerance and for minimal cost of resources. The code emerges at a coding transition when the mapping of codons to amino-acids becomes nonrandom. The emergence of the code is governed by the topology defined by the probable errors and is related to the map coloring problem.

Summary of Khorana's research

Ribonucleic acid (RNA) with two repeating units ($\text{UCUCUCU} \rightarrow \text{UCU}$ CUC UCU) produced two alternating amino acids. This, combined with the Nirenberg and Leder experiment, showed that UCU codes for Serine and CUC codes for Leucine. RNAs with three repeating units ($\text{UACUACUA} \rightarrow \text{UAC UAC UAC}$, or ACU ACU ACU, or CUA CUA CUA) produced three different strings of amino acids. RNAs with

four repeating units including UAG, UAA, or UGA, produced only dipeptides and tripeptides thus revealing that UAG, UAA and UGA are stop codons. With this, Khorana and his team had established that the mother of all codes, the biological language common to all living organisms, is spelled out in three-letter words: each set of three nucleotides codes for a specific amino acid. Their Nobel lecture was delivered on December 12, 1968. To do this Khorana was also the first to synthesize oligonucleotides, that is, strings of nucleotides.

The table of Genetic Code

		2nd base						
		T	C	A	G			
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine
	TTC	(Phe/F) Phenylalanine	TCC	(Ser/S) Serine	TAC	(Tyr/Y) Tyrosine	TGC	(Cys/C) Cysteine
	TTA	(Leu/L) Leucine	TCA	(Ser/S) Serine	TAA	Ochre (Stop)	TGA	Opal (Stop)
	TTG	(Leu/L) Leucine	TCG	(Ser/S) Serine	TAG	Amber (Stop)	TGG	(Trp/W) Tryptophan
	CTT	(Leu/L) Leucine	CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine
	CTC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
C	CTA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
	CTG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine
	ATC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
	ATA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
	ATG	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
A	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine
	GTC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
	GTA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
	GTG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

Degeneracy of the genetic code

Degeneracy is the redundancy of the genetic code. The genetic code has redundancy but no ambiguity (above for the full correlation). For example, although codons GAA and GAG both specify glutamic acid (redundancy), neither of them specifies any other amino acid (no ambiguity). The codons encoding one amino acid may differ in any of their three positions. For example the amino acid glutamic acid is specified by GAA and GAG codons (difference in the third position), the amino acid leucine is specified by UUA, UUG, CUU, CUC, CUA, CUG codons (difference in the first or third position), while the amino acid serine is specified by UCA, UCG, UCC, UCU, AGU, AGC (difference in the first, second or third position).

A position of a codon is said to be a fourfold degenerate site if any nucleotide at this position specifies the same amino acid. For example, the third position of the glycine codons (GGA, GGG, GGC, GGU) is a fourfold degenerate site, because all nucleotide substitutions at this site are synonymous; i.e., they do not change the amino acid. Only the third positions of some codons may be fourfold degenerate. A position of a codon is said to be a twofold degenerate site if only two of four possible nucleotides at this position specify the same amino acid. For example, the third position of the glutamic acid codons (GAA, GAG) is a twofold degenerate site. In twofold degenerate sites, the equivalent nucleotides are always either two purines (A/G) or two pyrimidines (C/U), so only transversional substitutions (purine to pyrimidine or pyrimidine to purine) in twofold degenerate sites are nonsynonymous.

A position of a codon is said to be a non-degenerate site if any mutation at this position results in amino acid substitution. There is only one threefold degenerate site where changing to three of the

four nucleotides may have no effect on the amino acid (depending on what it is changed to), while changing to the fourth possible nucleotide always results in an amino acid substitution. This is the third position of an isoleucine codon: AUU, AUC, or AUA all encode isoleucine, but AUG encodes methionine. In computation this position is often treated as a twofold degenerate site.

There are three amino acids encoded by six different codons: serine, leucine, and arginine. Only two amino acids are specified by a single codon. One of these is the amino-acid methionine, specified by the codon AUG, which also specifies the start of translation; the other is tryptophan, specified by the codon UGG. The degeneracy of the genetic code is what accounts for the existence of synonymous mutations.

Degeneracy results because there are more codons than encodable amino acids. For example, if there were two bases per codon, then only 16 amino acids could be coded for ($4^2=16$). Because at least 21 codes are required (20 amino acids plus stop), and the next largest number of bases is three, then 4^3 gives 64 possible codons, meaning that some degeneracy must exist.

These properties of the genetic code make it more fault-tolerant for point mutations. For example, in theory, fourfold degenerate codons can tolerate any point mutation at the third position, although codon usage bias restricts this in practice in many organisms; twofold degenerate codons can tolerate one out of the three possible point mutations at the third position. Since transition mutations (purine to purine or pyrimidine to pyrimidine mutations) are more likely than transversion (purine to pyrimidine or vice-versa) mutations, the equivalence of purines or that of pyrimidines at twofold degenerate sites adds a further fault-tolerance.

Despite the redundancy of the genetic code, single point mutations can still cause dysfunctional proteins. For example, a mutated hemoglobin gene causes sickle-cell disease. In the mutant hemoglobin a hydrophilic glutamate (Glu) is substituted by the hydrophobic valine (Val), that is, GAA or GAG becomes GUA or GUG. The substitution of glutamate by valine reduces the solubility

of Beta globulins| β -globin which causes hemoglobin to form linear polymers linked by the hydrophobic interaction between the valine groups, causing sickle-cell deformation of erythrocytes. Sickle-cell disease is generally not caused by a *de novo* mutation. Rather it is selected for in geographic regions where malaria is common (in a way similar to thalassemia), as heterozygous people have some resistance to the malarial *Plasmodium* parasite (heterozygote advantage).^[5]

These variable codes for amino acids are allowed because of modified bases in the first base of the anticodon of the tRNA, and the base-pair formed is called a wobble base pair. The modified bases include inosine and the Non-Watson-Crick U-G basepair.^[6]

Initiation and Termination codon

Initiation or Start Codon

The start codon is generally defined as the point, sequence, at which a ribosome begins to translate a sequence of RNA into amino acids. When an RNA transcript is “read” from the 5’ carbon to the 3’ carbon by the ribosome the start codon is the first codon on which the tRNA bound to Met, methionine, and ribosomal subunits attach. **ATG and AUG denote sequences of DNA and RNA, respectively, that are the start codon or initiation codon encoding the amino acid methionine (Met) in eukaryotes and a modified Met (fMet) in prokaryotes.** The principle called the Central dogma of molecular biology describes the process of translation of a gene to a protein. Specific sequences of DNA act as a template to synthesize mRNA in a process termed “transcription” in the nucleus. This mRNA is exported from the nucleus into the cytoplasm of the cell and acts as a template to synthesize protein in a process called “translation.” Three nucleotide bases specify one amino acid in the genetic code,

a mapping encoded in the tRNA of the organism. The first three bases of the coding sequence (CDS) of mRNA to be translated into protein are called a start codon or initiation codon. The start codon is almost always preceded by an untranslated region 5' UTR. The start codon is typically AUG (or ATG in DNA; this also encodes methionine). Very rarely in higher organisms (eukaryotes) are non AUG start codons used. In addition to AUG, alternative start codons, mainly **GUG** and **UUG** are used in prokaryotes. For example *E. coli* uses 83% ATG (AUG), 14% GTG (GUG), 3% TTG (UUG) and one or two others (e.g., ATT and CTG).

Termination or Stop codon

In the genetic code, a stop codon (also known as termination codon) is a nucleotide triplet within messenger RNA that signals a termination of translation. Proteins are based upon polypeptides, which are unique sequences of amino acids; and most codons in messenger RNA correspond to the addition of an amino acid to a growing polypeptide chain, which may ultimately become a protein – stop codons signal the termination of this process, releasing the amino acid chain.

Stop codons were historically given many different names, as they each corresponded to a distinct class of mutants that all behaved in a similar manner. **These mutants were first isolated within bacteriophages** (T4 and lambda), viruses that infect the bacteria *Escherichia coli*. Mutations in viral genes weakened their infectious

		Second letter					
		U	C	A	G		
First letter	U	UUU Phe UUC UUA UUG	UCU Ser UCG	UAU Tyr UAC UAA Stop UAG Stop	UGC Cys UGA Stop UGG Trp	UC CA G	
	C	CUU Leu CUC CUA CUG	CCU Pro CCC CCA CCG	CAU His CAC CAA Gin CAG	CGU Arg CGC CGA CGG	UC CA G	
A	AUU Ile AUC AAA AUG Met	ACU ACC ACA ACG	Thr	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	UC CA G	
G	GUU Val GUC GUA GUG	GCU Ala GCC GCA GCG	GAA Asp GAC GAA Glu GAG	GGU Gly GGC GGA GGG	UC CA G		

Genetic Code Chart

ability, sometimes creating viruses that were able to infect and grow within only certain varieties of E coli.

1. **Amber mutations** were the first set of nonsense mutations to be discovered. They were isolated by Richard Epstein and Charles Steinberg, but named after their friend Harris Bernstein (see Edgar pgs. 580-581^[7]) for the story behind this incident) Viruses with amber mutations are characterized by their ability to infect only certain strains of bacteria, known as amber suppressors. These bacteria carry their own mutation that allow a recovery of function in the mutant viruses. For example, a mutation in the tRNA that recognizes the amber stop codon allows translation to “read through” the codon and produce full-length protein, thereby recovering the normal form of the protein and “suppressing” the amber mutation. Thus, amber mutants are an entire class of virus mutants that can grow in bacteria that contain amber suppressor mutations.
2. **Ochre** Ochre mutation was the second stop codon mutation to be discovered. Given a color name to match the name of amber mutants, ochre mutant viruses had a similar property in that they recovered infectious ability within certain suppressor strains of bacteria. The set of ochre suppressors was distinct from amber suppressors, so ochre mutants were inferred to correspond to a different nucleotide triplet. Through a series of mutation experiments comparing these mutants with each other and other known amino acid codons, Sydney Brenner concluded that the amber and ochre mutations corresponded to the nucleotide triplets “UAG” and “UAA”.^[8]
3. Opal mutations or umber mutations the third and last stop codon in the standard genetic code was discovered soon after, corresponding to the nucleotide triplet “UGA”. Nonsense mutations that created this premature stop codon were later called opal mutations or umber mutations.

In RNA: UAG (“amber”) UAA (“ochre”) UGA (“opal”)

In DNA: TAG (“amber”) TAA (“ochre”) TGA (“opal” or “umber”).

Exceptions to the Universal Genetic Code (UGC) in mitochondria

Organism	Codon	Standard	Novel
Mammalian	AGA, AGG	Arginine	Stop codon
	AUA	Isoleucine	Methionine
	UGA	Stop codon	Tryptophan
Invertebrates	AGA, AGG	Arginine	Serine
	AUA	Isoleucine	Methionine
	UGA	Stop codon	Tryptophan
Yeast	AUA	Isoleucine	Methionine
	UGA	Stop codon	Tryptophan
	CUA	Leucine	Threonine

Facts to be remembered

Exceptions to the genetic code: Although the vast majority of living organisms today use the standard genetic code, geneticists have discovered a few variations on this code. Moreover, these variants are found in different evolutionary lineages and consist of different translations of a few codons.

The CUG codon, usually translated as leucine , corresponds to the serine 2 in many species of fungi Candida 3.

Many species of green algae of the genus Acetabularia use stop codons UAG and UAA to encode glycine.

Many ciliates like Paramecium tetraurelia , Tetrahymena thermophila or Stylopnychia 4 lemnae use codons UAG and UAA to code for glutamine instead of stop. UGA is the one stop codon used by these cells.

The ciliate Euplotes octocarinatus uses the codon UGA to encode cysteine, leaving UAG and UAA as stop signs.

In the three kingdoms of life , we sometimes find a twenty-first

amino acid, selenocysteine , encoded by the UGA codon (normally a stop codon).

In archaea and eubacteria , a twenty-second amino acid, pyrrolysine is sometimes met, encoded by UAG (also usually a stop codon).

The first amino acid incorporated (determined by the start codon AUG) is a methionine in most eukaryotes , more rarely a valine (in some eukaryotes), and formyl-methionine in most prokaryotes . In addition, this codon is GUG or GUU sometimes in some prokaryotes.

We therefore believe that life today originally had a smaller number of amino acids. These amino acids have been modified and have seen their numbers increase (by a phenomenon similar to the formation of sélenocytéine and pyrrolysine derived from serine and lysine, respectively, modified as they are on their transfer RNA on the ribosome .) These new amino acids were then used a subset of transfer RNAs and their associated coding. Maybe we notice signs of this phenomenon with glutamine , which in some bacteria, derived from glutamate still attached to its tRNA.

Another exception: the code is sometimes ambiguous. For example, the codon UGA is in the same organism (Escherichia coli , for example) sometimes code for the 21st amino acid mentioned above (selenocysteine) or “stop”.

References

1. ↑ Gamow, G. 1954. Possible relation between deoxyribonucleic acid and protein structure. *Nature* **173**: 318.
2. ↑ “General nature of the genetic code for proteins”. *Nature* **192**: 1227–32. 1961. PMID 13882203.
3. ↑ “The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides”. *Proc. Natl. Acad. Sci. U.S.A.* **47**: 1588–602. 1961. PMID 14479932.

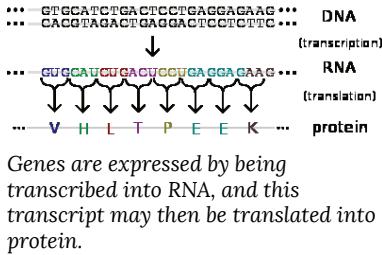
4. ↑ Genetic Code
5. ↑ Hebbel RP (2003). “Sickle hemoglobin instability: a mechanism for malarial protection”. *Redox Rep.* **8** (5): 238–40. doi:10.1179/135100003225002826. PMID 14962356.
6. ↑ Varani G, McClain WH (July 2000). “The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems”. *EMBO Rep.* **1** (1): 18–23. doi:10.1093/embo-reports/kvd001. PMID 11256617.
7. ↑ “The genome of bacteriophage T4: an archeological dig”. *Genetics* **168** (2): 575–82. 2004. PMID 15514035.
8. ↑ Brenner S. A Life in Science 2001 (see pgs. 101-104) Published by BioMed Central Limited ISBN 0954027809 ISBN 978-0954027803

I7.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as ribosomal RNA (rRNA) genes or transfer RNA (tRNA) genes, the product is a functional RNA. The process of gene expression is used by all known life – eukaryotes (including multicellular organisms), prokaryotes (bacteria and archaea) and viruses – to generate the macromolecular machinery for life. Several steps in the gene expression process may be modulated, including the transcription, RNA splicing, translation, and post-translational modification of a protein. Gene regulation gives the cell control over structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism. Gene regulation may also serve as a substrate for evolutionary change, since control of the timing, location, and amount of gene expression can have a profound effect on the functions (actions) of the gene in a cell or in a multicellular organism. In genetics, gene expression is the most fundamental level at which genotype gives rise to the phenotype. The genetic code stored in DNA in form of nucleotide sequence is “interpreted” by gene expression, and the properties of the expression products give rise to the organism’s phenotype.^[1] A molecule which allows the genetic material to be realized as a protein was first hypothesized by **François Jacob and Jacques Monod**. RNA synthesis by RNA polymerase was established in vitro by several laboratories by 1965; however, the RNA synthesized by these enzymes had properties that suggested the existence of an additional factor needed to terminate transcription correctly. In 1972, Walter Fiers became the first person to actually prove the existence of the terminating enzyme. Roger D. Kornberg won the 2006 Nobel Prize in Chemistry “for his studies of the molecular basis of eukaryotic transcription.”

Contents

- 1 Transcription
- 2 One gene-one enzyme hypothesis
 - 2.1 One gene-one polypeptide
- 3 Operon
 - 3.1 Structure of an operon
 - 3.1.1 Prokaryotic promoters
 - 3.1.2 Eukaryotic promoters
 - 3.2 Enhancer
 - 3.2.1 Corepressor
 - 3.3 Riboswitch
- 4 Lac operon
 - 4.1 Lac repressor (LacI)
- 5 Trp operon
- 6 Arabinose operon
- 7 Housekeeping gene
- 8 Regulation of gene expression
- 9 Tools for studying gene expression
 - 9.1 Vector
 - 9.2 PCR
 - 9.3 Restriction enzymes
 - 9.4 Cloning of gene and its expression
 - 9.5 Reporter gene
- 10 References



Transcription

Generation of RNA from DNA is known as transcription. In other

word transcription is the process of creating a complementary RNA copy of a sequence of DNA. During transcription, a DNA sequence is read by **RNA polymerase**, which produces a complementary, antiparallel RNA strand. As opposed to DNA replication, transcription results in an RNA complement that includes uracil (U) in all instances where thymine (T) would have occurred in a DNA complement.^[2] Transcription can be explained easily in 4 or 5 simple steps, each moving like a wave along the DNA.

As the Hydrogen Bonds Break DNA unwinds.

The free nucleotides of the RNA, pair with complementary DNA base RNA sugar-phosphate backbone forms. (by RNA Polymerase.)

Hydrogen bonds of the untwisted RNA-DNA "ladder" break, freeing The RNA is further processed and then moves through the small n

Transcription is the first step leading to gene expression.

The stretch of DNA transcribed into an RNA molecule is called a transcription unit and encodes at least one gene. If the gene transcribed encodes a protein, the result of transcription is messenger RNA (mRNA), which will then be used to create that protein via the process of translation. Alternatively, the transcribed gene may encode for either ribosomal RNA (rRNA) or transfer RNA (tRNA), other components of the protein-assembly process, or other ribozymes.

A DNA transcription unit encoding for a protein contains not only the sequence that will eventually be directly translated into the protein (the coding sequence) but also regulatory sequences that direct and regulate the synthesis of that protein. The regulatory sequence before (upstream from) the coding sequence is called the 5'UTR (five prime untranslated region), and the sequence following (downstream from) the coding sequence is called the 3'UTR (three prime untranslated region). Transcription has some proofreading mechanisms, but they are fewer and less effective than the controls for copying DNA; therefore, transcription has a lower copying fidelity than DNA replication. As in DNA replication, DNA is read from 3' → 5' during transcription. Meanwhile, the complementary

RNA is created from the $5' \rightarrow 3'$ direction. This means its 5' end is created first in base pairing. Although DNA is arranged as two antiparallel strands in a double helix, only one of the two DNA strands, called the template strand, is used for transcription. This is because RNA is only single-stranded, as opposed to double-stranded DNA. The other DNA strand is called the coding strand, because its sequence is the same as the newly created RNA transcript (except for the substitution of uracil for thymine). The use of only the $3' \rightarrow 5'$ strand eliminates the need for the Okazaki fragments seen in DNA replication. Transcription is divided into 5 stages: pre-initiation, initiation, promoter clearance, elongation and termination.

One gene-one enzyme hypothesis

The one gene-one enzyme hypothesis is the idea that genes act through the production of enzymes, with each gene responsible for producing a single enzyme that in turn affects a single step in a metabolic pathway. The concept was proposed by **George Beadle and Edward Tatum** in an influential 1941 paper on genetic mutations in the mold *Neurospora crassa*,^[3] and subsequently was dubbed the “one gene-one enzyme hypothesis” by their collaborator Norman Horowitz. It is often considered the first significant result in what came to be called molecular biology. Although it has been extremely influential, the hypothesis was recognized soon after its proposal to be an oversimplification. Even the subsequent reformulation of the “one gene-one polypeptide” hypothesis is now considered too simple to describe the relationship between genes and proteins.^[4]

What is *Neurospora*? *Neurospora crassa* is a type of red bread mold of the phylum **Ascomycota**. The genus

name, meaning “**nerve spore**” refers to the characteristic striations on the spores.

N. crassa is used as a model organism because it is easy to grow and has a haploid life cycle that makes genetic analysis simple since recessive traits will show up in the offspring. Analysis of genetic recombination is facilitated by the ordered arrangement of the products of meiosis in *Neurospora* ascospores. Its entire genome of seven chromosomes has been sequenced. *Neurospora* was used by Edward Tatum and George Wells Beadle in their experiments for which they won the Nobel Prize in Physiology or Medicine in 1958. Beadle and Tatum exposed *N. crassa* to x-rays, causing mutations. They then observed failures in metabolic pathways caused by errors in specific enzymes. This led them to propose the “one gene, one enzyme” hypothesis that specific genes code for specific proteins. Their hypothesis was later elaborated to enzyme pathways by **Norman Horowitz**, also working on *Neurospora*.

One gene-one polypeptide

By the early 1950s, advances in biochemical genetics—spurred in part by the original hypothesis—made the one gene-one enzyme hypothesis seem very unlikely (at least in its original form). Beginning in 1957, Vernon Ingram and others showed through protein fingerprinting that genetic variations in proteins (such as sickle cell hemoglobin) could be limited to differences in just a single polypeptide chain in a multimeric protein, leading to a “one gene-one polypeptide” hypothesis instead. According to geneticist

Rowland H. Davis, “By 1958 – indeed, even by 1948 – one gene, one enzyme was no longer a hypothesis to be resolutely defended; it was simply the name of a research program.” Presently, the one gene-one polypeptide perspective cannot account for the various spliced versions in many eukaryote organisms which use a spliceosome to individually prepare a RNA transcript depending on the various inter- and intra-cellular environmental signals. This splicing was discovered in 1977 by Phillip Sharp and Richard J. Roberts.

Operon

An operon is a functioning unit of genomic material containing a cluster of genes under the control of a single regulatory signal or promoter. The genes are transcribed together into an mRNA strand and either translated together in the cytoplasm, or undergo trans-splicing to create monocistronic mRNAs that are translated separately, i.e. several strands of mRNA that each encode a single gene product. The result of this is that the genes contained in the operon are either expressed together or not at all. Several genes must be both co-transcribed and co-regulated to define an operon. Originally operons were thought to exist solely in prokaryotes but since the discovery of the first operons in eukaryotes in the early 1990s, more evidence has arisen to suggest they are more common than previously assumed.

Operons occur primarily in prokaryotes but also in some eukaryotes, including nematodes such as *C. elegans*, and *Drosophila melanogaster* flies. rRNA genes often exist in operons that have been found in a range of eukaryotes including chordates. An operon is made up of several structural genes arranged under a common promoter and regulated by a common operator. It is defined as a set of adjacent structural genes, plus the adjacent regulatory signals that affect transcription of the structural genes. The regulators of a given operon, including repressors,

corepressors, and activators, are not necessarily coded for by that operon. The location and condition of the regulators, promoter, operator and structural DNA sequences can determine the effects of common mutations. Operons are related to regulons, stimulons and modulons. Whereas operons contain a set of genes regulated by the same operator, regulons contain a set of genes under regulation by a single regulatory protein, and stimulons contain a set of genes under regulation by a single cell stimulus.^[5]

Structure of an operon

Promoter – a nucleotide sequence that enables a gene to be transcribed. The promoter is recognized by RNA polymerase, which then initiates transcription. In RNA synthesis, promoters indicate which genes should be used for messenger RNA creation – and, by extension, control which proteins the cell manufactures.

Operator – a segment of DNA that a regulator binds to. It is classically defined in the lac operon as a segment between the promoter and the genes of the operon. In the case of a repressor, the repressor protein physically obstructs the RNA polymerase from transcribing the genes.

Structural genes – the genes that are co-regulated by the operon.

Prokaryotic promoters

In prokaryotes, the promoter consists of two short sequences at -10 and -35 positions upstream from the transcription start site. Sigma factors not only help in enhancing RNAP binding to the promoter but also help RNAP target specific genes to transcribe. The sequence at **-10 is called the Pribnow box**, or the -10 element, and usually consists of the six nucleotides TATAAT. The Pribnow box is absolutely essential to start transcription in prokaryotes.

The other sequence at -35 (the -35 element) usually consists of the seven nucleotides TTGACAT. Its presence allows a very high transcription rate. Both of the above consensus sequences, while conserved on average, are not found intact in most promoters. On average only 3 of the 6 base pairs in each consensus sequence is found in any given promoter. No promoter has been identified to date that has intact consensus sequences at both the -10 and -35; artificial promoters with complete conservation of the -10/-35 hexamers has been found to promote RNA chain initiation at very high efficiencies. Some promoters contain a UP element (consensus sequence 5'-AAAWWTWTTTNNNAAANN-3'; W = A or T; N = any base) centered at -50; the presence of the -35 element appears to be unimportant for transcription from the UP element-containing promoters. It should be noted that the above promoter sequences are only recognized by the sigma-70 protein that interacts with the prokaryotic RNA polymerase. Complexes of prokaryotic RNA polymerase with other sigma factors recognize totally different core promoter sequences.

<-- upstream
5' -XXGGGCCGGGTGGTTGGGCCGAAGGGTTGGCCG
-35 -10 Gene to be transcribed

Eukaryotic promoters

Eukaryotic promoters are extremely diverse and are difficult to characterize. They typically lie upstream of the gene and can have regulatory elements several kilobases away from the transcriptional start site(enancers). In eukaryotes, the transcriptional complex can cause the DNA to bend back on itself, which allows for placement of regulatory sequences far from the actual site of transcription. Many eukaryotic promoters, between 10 and 20% of all genes contain a TATA box (sequence TATAAA), which in turn binds a TATA binding protein which assists in the formation of the RNA polymerase

transcriptional complex. The TATA box typically lies very close to the transcriptional start site (often within 50 bases).

Eukaryotic promoter regulatory sequences typically bind proteins called transcription factors which are involved in the formation of the transcriptional complex. An example is the E-box (sequence CACGTG), which binds transcription factors in the basic-helix-loop-helix (bHLH) family (e.g. BMAL1-Clock, cMyc).

Enhancer

An enhancer is a short region of DNA that can be bound with proteins (namely, the trans-acting factors, much like a set of transcription factors) to enhance transcription levels of genes (hence the name) in a gene cluster. While enhancers are usually cis-acting, an enhancer does not need to be particularly close to the genes it acts on, and need not be located on the same chromosome.

In eukaryotic cells the structure of the chromatin complex of DNA is folded in a way that functionally mimics the supercoiled state characteristic of prokaryotic DNA, so that although the enhancer DNA is far from the gene in regard to the number of nucleotides, it is geometrically close to the promoter and gene. This allows it to interact with the general transcription factors and RNA polymerase II. An enhancer may be located upstream or downstream of the gene that it

regulates.

Furthermore, an enhancer does not need to be located near to the transcription initiation site to affect the transcription of a gene, as some have been found to bind several hundred thousand base pairs upstream or downstream of the start site. **Enhancers do not act on the promoter region itself, but are bound by activator proteins.** These activator proteins interact with the mediator complex, which recruits polymerase II and the general transcription factors which then begin transcribing the genes. Enhancers can

also be found within introns. An enhancer's orientation may even be reversed without affecting its function. Additionally, an enhancer may be excised and inserted elsewhere in the chromosome, and still affect gene transcription. That is the reason that intron polymorphisms are checked though they are not translated.

Corepressor

A corepressor is a protein that decreases gene expression by binding to a transcription factor which contains a DNA binding domain. The corepressor is unable to bind DNA by itself. The corepressor can repress transcriptional initiation by recruiting histone deacetylases which catalyze the removal of acetyl groups from lysine residues. This increases the positive charge on histones which strengthens in the interaction between the histones and DNA, making the latter less accessible to transcription.

Riboswitch

In molecular biology, a riboswitch is a part of an mRNA molecule that can directly bind a small target molecule, and whose binding of the target affects the gene's activity. Thus, *an mRNA that contains a riboswitch is directly involved in regulating its own activity*, in response to the concentrations of its target molecule. The discovery that modern organisms use RNA to bind small molecules, and discriminate against closely related analogs, significantly expanded the known natural capabilities of RNA beyond its ability to code for proteins or to bind other RNA or protein macromolecules. The original definition of the term "riboswitch" specified that they directly sense small-molecule metabolite concentrations. Although this definition remains in common use, some biologists have used a broader definition that includes other *cis*-regulatory RNAs.

However, this article will discuss only metabolite-binding riboswitches. Most known riboswitches occur in bacteria, but functional riboswitches of one type (the **TPP riboswitch**) have been discovered in plants and certain fungi. TPP riboswitches have also been predicted in archaea, but have not been experimentally tested.^[6]

Lac operon

The **lac operon** is an operon required for the transport and metabolism of lactose in *Escherichia coli* and some other enteric bacteria. It consists of three adjacent structural genes, **lacZ**, **lacY** and **lacA**. The lac operon is regulated by several factors including the availability of glucose and of lactose. Gene regulation of the lac operon was the first complex genetic regulatory mechanism to be elucidated and is one of the foremost examples of prokaryotic gene regulation.



In its natural environment, the lac operon allows for the effective digestion of lactose. The cell can use lactose as an energy source by producing the enzyme **β -galactosidase** to digest that lactose into glucose and galactose. However, it would be inefficient to produce enzymes when there is no lactose available, or if there is a more readily-available energy source available such as glucose. The lac operon uses a two-part control mechanism to ensure that the cell expends energy producing β -galactosidase, β -galactoside permease and thiogalactoside transacetylase (also known as galactoside O-acetyltransferase) only when necessary. It achieves this with the lac repressor, which halts production in the absence of lactose, and the Catabolite activator protein (CAP), which assists in production in the absence of glucose. This dual control mechanism causes the sequential utilization of glucose and lactose in two distinct

growth phases, known as diauxie. Similar diauxic growth patterns have been observed in bacterial growth on mixtures of other sugars as well, such as mixtures of glucose and xylose, or of glucose and arabinose, etc. The genetic control mechanisms underlying such diauxic growth patterns are known as *xyl* operon and *ara* operon, etc.^[7] The lac operon consists of three structural genes, and a promoter, a terminator, regulator, and an operator.

The three structural genes are: lacZ, lacY, and lacA.

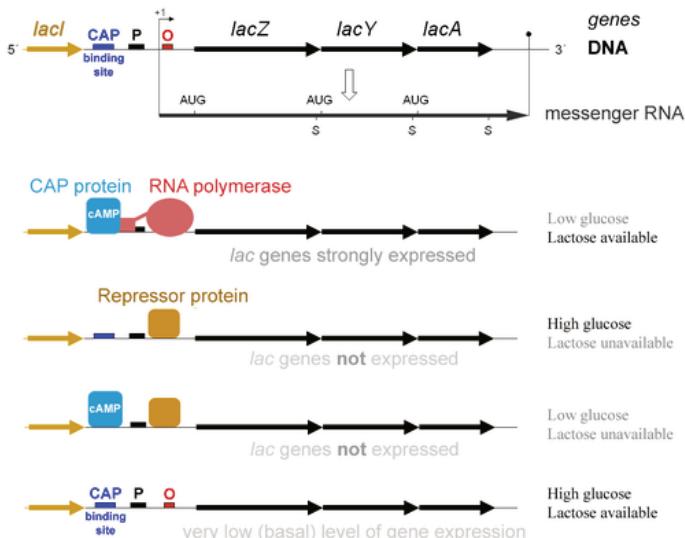
lacZ encodes β-galactosidase (LacZ), an intracellular enzyme that cleaves the disaccharide lactose into glucose and galactose.

lacY encodes β-galactoside permease (LacY), a membrane-bound transport protein that pumps lactose into the cell.

lacA encodes β-galactoside transacetylase (LacA), an enzyme that transfers an acetyl group from acetyl-CoA to β-galactosides.

Only lacZ and lacY appear to be necessary for lactose catabolism.

The lac Operon and its Control Elements



lac operon in detail

Lac repressor (*LacI*)

The lac repressor was first isolated by Walter Gilbert and Benno Müller-Hill in 1966. They were able to show, *in vitro*, that the protein bound to DNA containing the lac operon, and released the DNA when IPTG was added. (IPTG is an allolactose analog.) They were also able to isolate the portion of DNA bound by the protein by using the enzyme deoxyribonuclease, which breaks down DNA. After treatment of the repressor-DNA complex, some DNA remained, suggesting that it had been masked by the repressor. This was later confirmed. These experiments were important, as they confirmed the mechanism of the lac operon, earlier proposed by Jacques Monod and Francois Jacob. The structure of the lac repressor protein consists of three distinct regions:

a core region (which binds allolactose) a tetramerization region (which joins four monomers in an alpha-helix bundle) a DNA-binding region (in which two LacI proteins bind a single operator site) The lac repressor occurs as a tetramer (four identical subunits bound together). This can be viewed as two dimers, with each dimer being able to bind to a single lac operator. The two subunits each bind to a slightly separated (major groove) region of the operator. The promoter is slightly covered by the lac repressor so RNAP cannot bind to and transcribe the operon. The DNA binding region consists of a helix-turn-helix structural motif. Interactive, rotating 3D views of the repressor structure, some bound to DNA, including morphs of how it bends the DNA double helix, are available at Lac Repressor in Proteopedia. The lac repressor (*LacI*) operates by binding to the major groove of the operator region of the lac operon. This blocks RNA polymerase from binding, and so prevents transcription of the mRNA coding for the Lac proteins. When lactose is present, allolactose binds to the lac repressor, causing an allosteric change in its shape. In its changed state, the lac repressor is unable to bind to its cognate operator.

The lac gene and its derivatives are amenable to use as a reporter

gene in a number of bacterial-based selection techniques such as two hybrid analysis, in which the successful binding of a transcriptional activator to a specific promoter sequence must be determined. In LB plates containing X-gal, the colour change from white colonies to a shade of blue corresponds to about 20-100 β -galactosidase units, while tetrazolium lactose and MacConkey lactose media have a range of 100-1000 units, being most sensitive in the high and low parts of this range respectively. Since MacConkey lactose and tetrazolium lactose media both rely on the products of lactose breakdown, they require the presence of both lacZ and lacY genes. The many lac fusion techniques which include only the lacZ gene are thus suited to the X-gal plates or ONPG liquid broths.^[8]

Trp operon

Trp operon is an operon – a group of genes that are used, or transcribed, together – that codes for the components for production of tryptophan. The Trp operon is present in many bacteria, but was first characterized in *Escherichia coli*. It is regulated so that when tryptophan is present in the environment, it is not used. It was an important experimental system for learning about gene regulation, and is commonly used to teach gene regulation.

Discovered in 1953 by Jacques Monod and colleagues, the trp operon in *E. coli* was the first repressible operon to be discovered. While the lac operon can be activated by a chemical (allolactose), the tryptophan (Trp) operon is inhibited by a chemical (tryptophan). This operon contains five structural genes: **trp E, trp D, trp C, trp B, and trp A**, which encodes tryptophan synthetase. It also contains a promoter which binds to RNA polymerase and an operator which blocks transcription when bound to the protein synthesized by the repressor gene (trp R) that binds to the operator. In the lac operon,

allolactose binds to the repressor protein, allowing gene transcription, while in the trp operon, tryptophan binds to the repressor protein effectively blocking gene transcription. In both situations, repression is that of RNA polymerase transcribing the genes in the operon. Also unlike the lac operon, the trp operon contains a leader peptide and an attenuator sequence which allows for graded regulation.

It is an example of negative regulation of gene expression. Within the operon's regulatory sequence, the operator is blocked by the repressor protein in the presence of tryptophan (thereby preventing transcription) and is liberated in tryptophan's absence (thereby allowing transcription). The process of attenuation complements this regulatory action.^[9]

Arabinose operon

The L-arabinose operon of the model bacterium *Escherichia coli* has been a focus for research in molecular biology for over 40 years, and has been investigated extensively at the genetic, biochemical, physiological, and biophysical levels. It is controlled by a dual positive and negative system. There are 3 structural genes: **araB**, **araA**, and **araD**. They encode the metabolic enzymes for breaking down the monosaccharide sugar arabinose into D-xylulose-5-phosphate, which is then metabolised via the pentose phosphate pathway. The initiator region, containing an operator site as well as a promoter, is called araI (the last letter of araI is an uppercase letter "i"). Near this site lies the araC gene, which encodes a repressor protein. The AraC protein binds to initiator region araI.

Housekeeping gene

A housekeeping gene is typically a constitutive gene that is required for the maintenance of basic cellular function, and are found in all human cells. Although some housekeeping genes are expressed at relatively constant levels(such as HSP90 and Beta-actin), other housekeeping genes may vary depending on experimental conditions. The origin of the term “housekeeping gene” remains obscure. Literature from 1976 used the term to describe specifically tRNA and rRNA. Interpreting gene expression data can be problematic, with most human genes registering 5-10 copies per cell (possibly representing error). Housekeeping genes are expressed in at least 25 copies per cell and sometimes number in the thousands.

Regulation of gene expression

Regulation of gene expression refers to the control of the amount and timing of appearance of the functional product of a gene. Control of expression is vital to allow a cell to produce the gene products it needs when it needs them; in turn this gives cells the flexibility to adapt to a variable environment, external signals, damage to the cell, etc. Some simple examples of where gene expression is important are:

Control of Insulin expression so it gives a signal for blood glucose regulation

X chromosome inactivation in female mammals to prevent an “overdose” of the genes it contains.

Cyclin expression levels control progression through the eukaryotic cell cycle

More generally gene regulation gives the cell control over all structure and function, and is the basis for cellular differentiation,

morphogenesis and the versatility and adaptability of any organism. Any step of gene expression may be modulated, from the DNA–RNA transcription step to post-translational modification of a protein. The stability of the final gene product, whether it is RNA or protein, also contributes to the expression level of the gene – an unstable product results in a low expression level. In general gene expression is regulated through changes in the number and type of interactions between molecules that collectively influence transcription of DNA and translation of RNA. Numerous terms are used to describe types of genes depending on how they are regulated, these include: A constitutive gene is a gene that is transcribed continually compared to a facultative gene which is only transcribed when needed. A housekeeping gene is typically a constitutive gene that is transcribed at a relatively constant level. The housekeeping gene's products are typically needed for maintenance of the cell. It is generally assumed that their expression is unaffected by experimental conditions. Examples include actin, GAPDH and ubiquitin. A facultative gene is a gene which is only transcribed when needed compared to a constitutive gene. An inducible gene is a gene whose expression is either responsive to environmental change or dependent on the position in the cell cycle.^[10]

Transcriptional regulation

Regulation of transcription can be broken down into three main routes of influence; genetic (direct interaction of a control factor with the gene), modulation (interaction of a control factor with the transcription machinery) and epigenetic (non-sequence changes in DNA structure which influence transcription).

The lambda repressor transcription factor (green) binds as a dimer to major groove of DNA target (red and blue) and disables initiation of transcription. From PDB 1LMB. Direct interaction with DNA is the simplest and the most direct method by which a protein can change transcription levels. Genes often have several protein binding sites around the coding region with the specific function of regulating transcription. There are many classes of regulatory DNA binding sites known as enhancers, insulators, repressors and

silencers. The mechanisms for regulating transcription are very varied, from blocking key binding sites on the DNA for RNA polymerase to acting as an activator and promoting transcription by assisting RNA polymerase binding. The activity of transcription factors is further modulated by intracellular signals causing protein post-translational modification including phosphorylated, acetylated, or glycosylated. These changes influence a transcription factor's ability to bind, directly or indirectly, to promoter DNA, to recruit RNA polymerase, or to favor elongation of a newly synthesized RNA molecule. The nuclear membrane in eukaryotes allows further regulation of transcription factors by the duration of their presence in the nucleus which is regulated by reversible changes in their structure and by binding of other proteins. Environmental stimuli or endocrine signals may cause modification of regulatory proteins eliciting cascades of intracellular signals, which result in regulation of gene expression. More recently it has become apparent that there is a huge influence of non-DNA-sequence specific effects on translation. These effects are referred to as epigenetic and involve the higher order structure of DNA, non-sequence specific DNA binding proteins and chemical modification of DNA. In general epigenetic effects alter the accessibility of DNA to proteins and so modulate transcription.

In eukaryotes, DNA is organized in form of nucleosomes. Note how the DNA (blue and green) is tightly wrapped around the protein core made of histone octamer (ribbon coils), restricting access to the DNA. From PDB 1KX5. DNA methylation is a widespread mechanism for epigenetic influence on gene expression and is seen in bacteria and eukaryotes and has roles in heritable transcription silencing and transcription regulation. In eukaryotes the structure of chromatin, controlled by the histone code, regulates access to DNA with significant impacts on the expression of genes in euchromatin and heterochromatin areas.

Post-transcriptional regulation

In eukaryotes, where export of RNA is required before translation is possible, nuclear export is thought to provide additional control

over gene expression. All transport in and out of the nucleus is via the nuclear pore and transport is controlled by a wide range of importin and exportin proteins. Expression of a gene coding for a protein is only possible if the messenger RNA carrying the code survives long enough to be translated. In a typical cell an RNA molecule is only stable if specifically protected from degradation. RNA degradation has particular importance in regulation of expression in eukaryotic cells where mRNA has to travel significant distances before being translated. In eukaryotes RNA is stabilised by certain post-transcriptional modifications, particularly the 5' cap and polyadenylated tail. Intentional degradation of mRNA is used not just as a defence mechanism from foreign RNA (normally from viruses) but also as a route of mRNA destabilisation. If an mRNA molecule has a complementary sequence to a small interfering RNA then it is targeted for destruction via the RNA interference pathway.

Translational regulation

Neomycin is an example of a small molecule which reduces expression of all protein genes inevitably leading to cell death, thus acts as an antibiotic.

Direct regulation of translation is less prevalent than control of transcription or mRNA stability but is occasionally used. Inhibition of protein translation is a major target for toxins and antibiotics in order to kill a cell by overriding its normal gene expression control. Protein synthesis inhibitors include the antibiotic neomycin and the toxin ricin.

Protein degradation

Once protein synthesis is complete the level of expression of that protein can be reduced by protein degradation. There are major protein degradation pathways in all prokaryotes and eukaryotes of which the proteasome is a common component. An unneeded or damaged protein is often labelled for degradation by addition of ubiquitin.

Tools for studying gene expression

Vector

Plasmids used in genetic engineering are called vectors. Plasmids serve as important tools in genetics and biotechnology labs, where they are commonly used to multiply (make many copies of) or express particular genes. Many plasmids are commercially available for such uses. The gene to be replicated is inserted into copies of a plasmid containing genes that make cells resistant to particular antibiotics and a multiple cloning site (MCS, or polylinker), which is a short region containing several commonly used restriction sites allowing the easy insertion of DNA fragments at this location. Next, the plasmids are inserted into bacteria by a process called transformation. Then, the bacteria are exposed to the particular antibiotics. Only bacteria which take up copies of the plasmid survive, since the plasmid makes them resistant. In particular, the protecting genes are expressed (used to make a protein) and the expressed protein breaks down the antibiotics. In this way the antibiotics act as a filter to select only the modified bacteria. Now these bacteria can be grown in large amounts, harvested and lysed (often using the alkaline lysis method) to isolate the plasmid of interest. Another major use of plasmids is to make large amounts of proteins. In this case, researchers grow bacteria containing a plasmid harboring the gene of interest. Just as the bacteria produces proteins to confer its antibiotic resistance, it can also be induced to produce large amounts of proteins from the inserted gene. This is a cheap and easy way of mass-producing a gene or the protein it then codes for, for example, insulin or even antibiotics. However, a plasmid can only contain inserts of about 1–10 kbp. To clone longer lengths of DNA, lambda phage with lysogeny genes deleted, cosmids, bacterial artificial chromosomes or yeast artificial chromosomes could be used.^[11]

Modern vectors may encompass additional features besides the transgene insert and a backbone: Promoter: Necessary component for all vectors: used to drive transcription of the vector's transgene.

Genetic markers: Genetic markers for viral vectors allow for confirmation that the vector has integrated with the host genomic DNA.

Antibiotic resistance: Vectors with antibiotic-resistance open reading frames allow for survival of cells that have taken up the vector in growth media containing antibiotics through antibiotic selection.

Epitope: Vector contains a sequence for a specific epitope that is incorporated into the expressed protein. Allows for antibody identification of cells expressing the target protein.

β -galactosidase: Some vectors contain a sequence for β -galactosidase, an enzyme that digests galactose, within which a multiple cloning site, the region in which a gene may be inserted, is located. An insert successfully ligated into the vector will disrupt the β -galactosidase gene and disable galactose digestion. Cells containing vector with an insert may be identified using blue/white selection by growing cells in media containing an analogue of galactose (X-gal). Cells expressing β -galactosidase (therefore doesn't contain an insert) appear as blue colonies. White colonies would be selected as those that may contain an insert. Other proteins which may function similarly as a reporter include green fluorescent protein and luciferase.

Targeting sequence: Expression vectors may include encoding for a targeting sequence in the finished protein that directs the expressed protein to a specific organelle in the cell or specific location such as the periplasmic space of bacteria.

Protein purification tags: Some expression vectors include proteins or peptide sequences that allows for easier purification of the expressed protein. Examples include polyhistidine-tag, glutathione-S-transferase, and maltose binding protein. Some of these tags may also allow for increased solubility of the target protein. The target protein is fused to the protein tag, but a protease

cleavage site positioned in the polypeptide linker region between the protein and the tag allows the tag to be removed later.

Cosmids

Cosmids are predominantly plasmids with a bacterial oriV, an antibiotic selection marker and a cloning site, but they carry one, or more recently two cos sites derived from bacteriophage lambda. Depending on the particular aim of the experiment broad host range cosmids, shuttle cosmids or 'mammalian' cosmids (linked to SV40 oriV and mammalian selection markers) are available. The loading capacity of cosmids varies depending on the size of the vector itself but usually lies around 40–45 kb. The cloning procedure involves the generation of two vector arms which are then joined to the foreign DNA. Selection against wildtype cosmid DNA is simply done via size exclusion. Cosmids, however, always form colonies and not plaques. Also the clone density is much lower with around $10^5 - 10^6$ CFU per μg of ligated DNA. After the construction of recombinant lambda or cosmid libraries the total DNA is transferred into an appropriate *E.coli* host via a technique called *in vitro* packaging. The necessary packaging extracts are derived from *E.coli* cl857 lysogens (red- gam- Sam and Dam (head assembly) and Eam (tail assembly) respectively). These extracts will recognize and package the recombinant molecules *in vitro*, generating either mature phage particles (lambda-based vectors) or recombinant plasmids contained in phage shells (cosmids). These differences are reflected in the different infection frequencies seen in favour of lambda-replacement vectors. This compensates for their slightly lower loading capacity. Phage library are also stored and screened easier than cosmid (colonies!) libraries. Target DNA: the genomic DNA to be cloned has to be cut into the appropriate size range of restriction fragments. This is usually done by partial restriction followed by either size fractionation or dephosphorylation (using calf-intestine phosphatase) to avoid chromosome scrambling, i.e. the ligation of physically unlinked fragments.

Fosmids

Fosmids are similar to cosmids but are based on the bacterial F-

plasmid. The cloning vector is limited, as a host (usually *E. coli*) can only contain one fosmid molecule. Fosmids are 40 kb of random genomic DNA. Fosmid library is prepared from a genome of the target organism and cloned into a fosmid vector. Low copy number offers higher stability than comparable high copy number cosmids. Fosmid system may be useful for constructing stable libraries from complex genomes. Fosmid clones were used to help assess the accuracy of the Public Human Genome Sequence.

Bacterial artificial chromosome (BAC)

A bacterial artificial chromosome (BAC) is a DNA construct, based on a functional fertility plasmid (or F-plasmid), used for transforming and cloning in bacteria, usually *E. coli*. F-plasmids play a crucial role because they contain partition genes that promote the even distribution of plasmids after bacterial cell division. The bacterial artificial chromosome's usual insert size is 150–350 kbp, but can be greater than 700 kbp. A similar cloning vector called a PAC has also been produced from the bacterial P1-plasmid. BACs are often used to sequence the genome of organisms in genome projects, for example the Human Genome Project. A short piece of the organism's DNA is amplified as an insert in BACs, and then sequenced. Finally, the sequenced parts are rearranged *in silico*, resulting in the genomic sequence of the organism.

Yeast artificial chromosome (YAC)

A yeast artificial chromosome (YAC) is a vector used to clone DNA fragments larger than 100 kb and up to 3000 kb. YACs are useful for the physical mapping of complex genomes and for the cloning of large genes. First described in 1983 by Murray and Szostak, a YAC is an artificially constructed chromosome and contains the telomeric, centromeric, and replication origin sequences needed for replication and preservation in yeast cells. A YAC is built using an initial circular plasmid, which is typically broken into two linear molecules using restriction enzymes; DNA ligase is then used to ligate a sequence or gene of interest between the two linear molecules, forming a single large linear piece of DNA.[citation needed] Yeast expression vectors, such as YACs, YIps (yeast

integrating plasmids), and YEps (yeast episomal plasmids), have an advantage over bacterial artificial chromosomes (BACs) in that they can be used to express eukaryotic proteins that require posttranslational modification. However, YACs have been found to be less stable than BACs, producing chimeric effects.

Types of viral vectors

Retroviruses

Retroviruses are one of the mainstays of current gene therapy approaches. The recombinant retroviruses such as the Moloney murine leukemia virus have the ability to integrate into the host genome in a stable fashion. They contain a reverse transcriptase that allows integration into the host genome. They have been used in a number of FDA-approved clinical trials such as the SCID-X1 trial. Retroviral vectors can either be replication-competent or replication-defective. Replication-defective vectors are the most common choice in studies because the viruses have had the coding regions for the genes necessary for additional rounds of virion replication and packaging replaced with other genes, or deleted. These virus are capable of infecting their target cells and delivering their viral payload, but then fail to continue the typical lytic pathway that leads to cell lysis and death. Conversely, replication-competent viral vectors contain all necessary genes for virion synthesis, and continue to propagate themselves once infection occurs. Because the viral genome for these vectors is much lengthier, the length of the actual inserted gene of interest is limited compared to the possible length of the insert for replication-defective vectors. Depending on the viral vector, the typical maximum length of an allowable DNA insert in a replication-defective viral vector is usually about 8–10 kB. While this limits the introduction of many genomic sequences, most cDNA sequences can still be accommodated. The primary drawback to use of retroviruses such as the Moloney retrovirus involves the requirement for cells to be actively dividing for transduction. As a result, cells such as neurons are very resistant to infection and transduction by retroviruses. There is concern that

insertional mutagenesis due to integration into the host genome might lead to cancer or leukemia.

Lentiviruses

Lentiviruses are a subclass of Retroviruses. They have recently been adapted as gene delivery vehicles (vectors) thanks to their ability to integrate into the genome of non-dividing cells, which is the unique feature of Lentiviruses as other Retroviruses can infect only dividing cells. The viral genome in the form of RNA is reverse-transcribed when the virus enters the cell to produce DNA, which is then inserted into the genome at a random position by the viral integrase enzyme. The vector, now called a provirus, remains in the genome and is passed on to the progeny of the cell when it divides. The site of integration is unpredictable, which can pose a problem. The provirus can disturb the function of cellular genes and lead to activation of oncogenes promoting the development of cancer, which raises concerns for possible applications of lentiviruses in gene therapy. However, studies have shown that lentivirus vectors have a lower tendency to integrate in places that potentially cause cancer than gamma-retroviral vectors. More specifically, one study found that lentiviral vectors did not cause either an increase in tumor incidence or an earlier onset of tumors in a mouse strain with a much higher incidence of tumors. Moreover, clinical trials that utilized lentiviral vectors to deliver gene therapy for the treatment of HIV experienced no increase in mutagenic or oncologic events. For safety reasons lentiviral vectors never carry the genes required for their replication. To produce a lentivirus, several plasmids are transfected into a so-called packaging cell line, commonly HEK 293. One or more plasmids, generally referred to as packaging plasmids, encode the virion proteins, such as the capsid and the reverse transcriptase. Another plasmid contains the genetic material to be delivered by the vector. It is transcribed to produce the single-stranded RNA viral genome and is marked by the presence of the ψ (psi) sequence. This sequence is used to package the genome into the virion.

Adenoviruses

As opposed to lentiviruses, adenoviral DNA does not integrate into the genome and is not replicated during cell division. This limits their use in basic research, although adenoviral vectors are occasionally used in *in vitro* experiments. Their primary applications are in gene therapy and vaccination. Since humans commonly come in contact with adenoviruses, which cause respiratory, gastrointestinal and eye infections, they trigger a rapid immune response with potentially dangerous consequences. To overcome this problem scientists are currently investigating adenoviruses to which humans do not have immunity.

Adeno-associated viruses

Adeno-associated virus (AAV) is a small virus that infects humans and some other primate species. AAV is not currently known to cause disease and consequently the virus causes a very mild immune response. AAV can infect both dividing and non-dividing cells and may incorporate its genome into that of the host cell. These features make AAV a very attractive candidate for creating viral vectors for gene therapy.

PCR

PCR is used to amplify a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of up to ~10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size. A basic PCR set up requires several components and reagents. These components include:

DNA template that contains the DNA region (target) to be amplified.

Two primers that are complementary to the 3' (three prime) ends of each of the sense and anti-sense strand of the DNA target. Taq polymerase or another DNA polymerase with a temperature optimum at around 70 °C. Deoxynucleotide triphosphates (dNTPs), the building-blocks from which the DNA polymerase synthesizes a new DNA strand. Buffer solution, providing a suitable chemical environment for optimum activity and stability of the DNA polymerase. Divalent cations, magnesium or

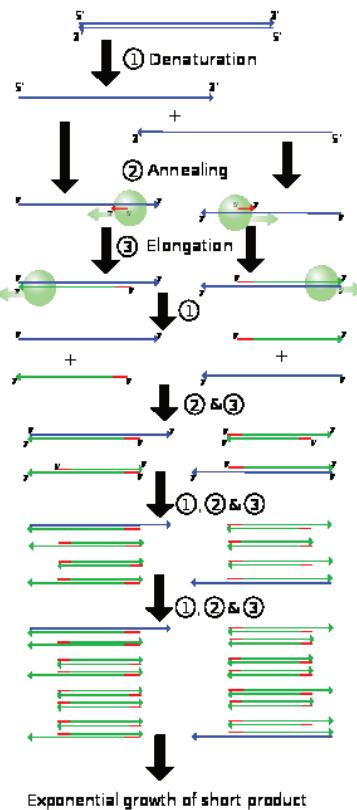


Figure 1: Schematic drawing of the PCR cycle. (1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C.
Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give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

manganese ions; generally Mg²⁺ is used, but Mn²⁺ can be utilized for PCR-mediated DNA mutagenesis, as higher Mn²⁺ concentration increases the error rate during DNA synthesis Monovalent cation potassium ions. The PCR is commonly carried out in a reaction volume of 10–200 µl in small reaction tubes (0.2–0.5 ml volumes) in a thermal cycler. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction (see below). Many modern thermal cyclers make use of the Peltier effect, which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled reaction tubes permit favorable thermal conductivity to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermocyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.^[12]

Procedure

Figure 1: Schematic drawing of the PCR cycle. (1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C. Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give shorter DNA products (green lines), which themselves are used as templates as PCR progresses. Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of 2-3 discrete temperature steps, usually three . The cycling is often preceded by a single temperature step (called hold) at a high temperature (>90 °C), and followed by one hold at the end for final product extension or brief storage. The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis, the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (T_m) of the primers.Initialization step: This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes. It is only

required for DNA polymerases that require heat activation by hot-start PCR. Denaturation step: This step is the first regular cycling event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules. Annealing step: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically the annealing temperature is about 3–5 degrees Celsius below the T_m of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis. Extension/elongation step: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to be amplified. As a rule-of-thumb, at its optimum temperature, the DNA polymerase will polymerize a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment. Final elongation: This single step is occasionally performed at a temperature of 70–74 °C for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended.^[13] Final hold: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.

To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplicon or amplicon),

agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products.

Restriction enzymes

Restriction enzymes recognize a specific sequence of nucleotides and produce a double-stranded cut in the DNA. While recognition sequences vary between 4 and 8 nucleotides, many of them are palindromic, which correspond to nitrogenous base sequences that read the same backwards and forwards.^[14] In theory, there are two types of palindromic sequences that can be possible in DNA. The mirror-like palindrome is similar to those found in ordinary text, in which a sequence reads the same forward and backwards on the same DNA strand (i.e., single stranded) as in GTAATG. The inverted repeat palindrome is also a sequence that reads the same forward and backwards, but the forward and backward sequences are found in complementary DNA strands (i.e., double stranded) as in GTATAC (Notice that GTATAC is complementary to CATATG).^[15] The inverted repeat is more common and has greater biological importance than the mirror-like.

EcoRI digestion produces “sticky” ends,



5'-GTATAAC-3'

:::::

3'-CATATG-5'

whereas SmaI restriction enzyme cleavage produces “blunt” ends



Recognition sequences in DNA differ for each restriction enzyme, producing differences in the length, sequence and strand orientation (5' end or the 3' end) of a sticky-end “overhang” of an enzyme restriction.^[16]

A palindromic recognition site reads the same on the reverse strand as it does on the forward strand

Different restriction enzymes that recognize the same sequence are known as neoschizomers. These often cleave in a different locales of the sequence; however, different enzymes that recognize and cleave in the same location are known as an isoschizomer.

Classification of Restriction Enzymes

Restriction endonucleases are categorized into three or four general groups (Types I, II and III) based on their composition and enzyme cofactor requirements, the nature of their target sequence, and the position of their DNA cleavage site relative to the target sequence. There are four classes of restriction endonucleases: types I, II, III and IV. All types of enzymes recognise specific short DNA sequences and carry out the endonucleolytic cleavage of DNA to give specific double-stranded fragments with terminal 5'-phosphates. They differ in their recognition sequence, subunit composition, cleavage position, and cofactor requirements, as summarised below:

Type I enzymes (EC 3.1.21.3) cleave at sites remote from recognition site; require both ATP and S-adenosyl-L-methionine to function; multifunctional protein with both restriction and methylase (EC 2.1.1.72) activities.

Type II enzymes (EC 3.1.21.4) cleave within or at short specific

distances from recognition site; most require magnesium; single function (restriction) enzymes independent of methylase.

Type III enzymes (EC 3.1.21.5) cleave at sites a short distance from recognition site; require ATP (but doesn't hydrolyse it); S-adenosyl-L-methionine stimulates reaction but is not required; exist as part of a complex with a modification methylase (EC 2.1.1.72). Type IV enzymes target methylated DNA.

Examples of restriction enzymes include:^[17]

Enzyme	Source	Recognition Sequence	Cut	
EcoRI	<i>Escherichia coli</i>	5' GAATTC 3' CTTAAG	5' ---G 3' ---CTTAA	AATTC---3' G---5'
EcoRII	<i>Escherichia coli</i>	5' CCWGG 3' GGWCC	5' --- 3' ---GGWCC	CCWGG---3' ---5'
BamHI	<i>Bacillus amyloliquefaciens</i>	5' GGATCC 3' CCTAGG	5' ---G 3' ---CCTAG	GATCC---3' G---5'
HindIII	<i>Haemophilus influenzae</i>	5' AAGCTT 3' TTCGAA	5' ---A 3' ---TTCGA	AGCTT---3' A---5'
TaqI	<i>Thermus aquaticus</i>	5' TCGA 3' AGCT	5' ---T 3' ---AGC	CGA---3' T---5'
NotI	<i>Nocardia otitidis</i>	5' GCGGCCGC 3' CGCCGGCG	5' ---GC 3' ---CGCCGG	GGCCGC---3' CG---5'
HinfI	<i>Haemophilus influenzae</i>	5' GANTCA 3' CTNAGT	5' ---G 3' ---CTNA	ANTC---3' G---5'
Sau3A	<i>Staphylococcus aureus</i>	5' GATC 3' CTAG	5' --- 3' ---CTAG	GATC---3' ---5'

Enzyme	Source	Recognition Sequence	Cut	
PovII*	<i>Proteus vulgaris</i>	5' CAGCTG 3' GTCGAC	5' ---CAG 3' ---GTC	CTG---3' GAC---5'
SmaI*	<i>Serratia marcescens</i>	5' CCCGGG 3' GGGCCC	5' ---CCC 3' ---GGG	GGG---3' CCC---5'
HaeIII*	<i>Haemophilus aegyptius</i>	5' GGCC 3' CCGG	5' ---GG 3' ---CC	CC---3' GG---5'
Hgal ^[18]	<i>Haemophilus gallinarum</i>	5' GACGC 3' CTGCG	5' ---NN 3' ---NN	NN---3' NN---5'
AluI*	<i>Arthrobacter luteus</i>	5' AGCT 3' TCGA	5' ---AG 3' ---TC	CT---3' GA---5'
EcoRV*	<i>Escherichia coli</i>	5' GATATC 3' CTATAG	5' ---GAT 3' ---CTA	ATC---3' TAG---5'
EcoP15I	<i>Escherichia coli</i>	5' CAGCAGN 25NN 3' GTCGTCN ₂₅ NN	5' ---CAGCAGN 25NN ---3' 3' ---GTCGTCN ₂₅ NN	NN---5'
KpnI	<i>Klebsiella pneumoniae</i>	5' GGTACC 3' CCATGG	5' ---GGTAC 3' ---C	C---3' CATGG---5'

Enzyme	Source	Recognition Sequence	Cut
PstI	<i>Providencia stuartii</i>	5' CTGCAG 3' GACGTC	5'---CTGCA G---3' 3'---G ACGTC---5'
SacI	<i>Streptomyces achromogenes</i>	5' GAGCTC 3' CTCGAG	5'---GAGCT C---3' 3'---C TCGAG---5'
SalI	<i>Streptomyces albus</i>	5' GTCGAC 3' CAGCTG	5'---G TCGAC---3' 3'---CAGCT G---5'
ScalI	<i>Streptomyces caespitosus</i>	5' AGTACT 3' TCATGA	5'---AGT ACT---3' 3'---TCA TGA---5'
SpeI	<i>Sphaerotilus natans</i>	5' ACTAGT 3' TGATCA	5'---A CTAGT---3' 3'---TGATC A---5'
SphI	<i>Streptomyces phaeochromogenes</i>	5' GCATGC 3' CGTAGC	5'---GCATG C---3' 3'---C GTACG---5'
StuI ^{[19][20]}	<i>Streptomyces tubercidicus</i>	5' AGGCCT 3' TCCGGA	5'---AGG CCT---3' 3'---TCC GGA---5'
XbaI	<i>Xanthomonas badrii</i>	5' TCTAGA 3' AGATCT	5'---T CTAGA---3' 3'---AGATC T---5'

Key:

* = blunt ends

N = C or G or T or A

W = A or T

Cloning of gene and its expression

In molecular biology Cloning refers to the procedure of isolating a defined DNA sequence and obtaining multiple copies of it *in vitro*. Cloning is frequently employed to amplify DNA fragments containing genes, but it can be used to amplify any DNA sequence such as promoters, non-coding sequences, chemically synthesised oligonucleotides and randomly fragmented DNA. Cloning is used in a wide array of biological experiments and technological applications such as large scale protein production and expression of gene in cell lines like HeLa cells.

In essence, in order to amplify any DNA sequence *in vivo* and *in vitro*, the sequence in question must be linked to primary sequence elements capable of directing the replication and propagation of themselves and the linked sequence in the desired target host. The required sequence elements differ according to host, but invariably include an origin of replication, and a selectable marker. In practice, however, a number of other features are desired and a variety of specialized cloning vectors exist that allow protein expression, tagging, single stranded RNA and DNA production and a host of other manipulations that are useful in downstream applications.

Recombinase-based cloning

A novel procedure of cloning or subcloning of any DNA fragment by inserting the special DNA fragment of interest into a special area of target DNA through interchange of the relevant DNA fragments.^[21]

This is a one-step reaction: simple, efficient, facilitating high throughput or automatic cloning and/or subcloning.^[22]

Restriction/ligation cloning

In the classical restriction and ligation cloning protocols, cloning of any DNA fragment essentially involves four steps: DNA

fragmentation with restriction endonucleases, ligation of DNA fragments to a vector, transfection, and screening/selection. Although these steps are invariable among cloning procedures a number of alternative routes can be selected at various points depending on the particular application; these are summarized as a 'cloning strategy'.

Isolation of insert

Initially, the DNA fragment to be cloned needs to be isolated. Preparation of DNA fragments for cloning can be accomplished in a number of alternative ways. Insert preparation is frequently achieved by means of polymerase chain reaction, but it may also be accomplished by restriction enzyme digestion, DNA sonication and fractionation by agarose gel electrophoresis. Chemically synthesized oligonucleotides can also be used if the target sequence size does not exceed the limit of chemical synthesis. Isolation of insert can be done by using shotgun cloning, c-DNA clones, gene machines (artificial chemical synthesis).

Transformation

Following ligation, the ligation product (plasmid) is transformed into bacteria for propagation. The bacteria is then plated on selective agar to select for bacteria that have the plasmid of interest. Individual colonies are picked and tested for the wanted insert. Maxiprep can be done to obtain large quantity of the plasmid containing the inserted gene.

Transfection

Following ligation, a portion of the ligation reaction, including vector with insert in the desired orientation is transfected into cells. A number of alternative techniques are available, such as chemical sensitization of cells, electroporation and biolistics. Chemical sensitization of cells is frequently employed since this does not require specialized equipment and provides relatively high transformation efficiencies. Electroporation is used when extremely high transformation efficiencies are required, as in very inefficient cloning strategies. Biolistics are mainly utilized in plant cell transformations, where the cell wall is a major obstacle in DNA

uptake by cells. The bacterial transformation is generally observed by blue white screening.

Selection

Finally, the transfected cells are cultured. As the aforementioned procedures are of particularly low efficiency, there is a need to identify the cells that contain the desired insert at the appropriate orientation and isolate these from those not successfully transformed. Modern cloning vectors include selectable markers (most frequently antibiotic resistance markers) that allow only cells in which the vector, but not necessarily the insert, has been transfected to grow. Additionally, the cloning vectors may contain colour selection markers which provide blue/white screening (via α -factor complementation) on X-gal medium. Nevertheless, these selection steps do not absolutely guarantee that the DNA insert is present in the cells. Further investigation of the resulting colonies is required to confirm that cloning was successful. This may be accomplished by means of PCR, restriction fragment analysis and/or DNA sequencing.

Genetic engineering

Genetic engineering is a method of changing the inherited characteristics of an organism in a predetermined way by altering its genetic material. This is often done to enable micro-organisms, such as bacteria or viruses, to synthesize increased yields of compounds, to form entirely new compounds, or to adapt to different environments. Other uses of this technology, which is also called recombinant DNA technology, include gene therapy, which is the supply of a functional gene to a person with a genetic disorder or with other diseases such as acquired immune deficiency syndrome (AIDS) or cancer, and the cloning of whole organisms.

Genetic engineering involves the manipulation of deoxyribonucleic acid, or DNA. Important tools in this process are restriction endonucleases (so-called restriction enzymes) that are produced by various species of bacteria. Restriction enzymes can recognize a particular sequence of the chain of chemical units, called nucleotide bases, which make up the DNA molecule and cut

the DNA at that location. Fragments of DNA generated in this way can be joined using other enzymes called ligases. Restriction enzymes and ligases therefore allow the specific cutting and reassembling of portions of DNA. Also important in the manipulation of DNA are so-called vectors, which are pieces of DNA that can self-replicate (produce copies of themselves) independently of the DNA in the host cell in which they are grown. Examples of vectors include plasmids, viruses, and artificial chromosomes. Vectors permit the generation of multiple copies of a particular piece of DNA, making this a useful method for generating sufficient quantities of material with which to work. The process of engineering a DNA fragment into a vector is called “molecular cloning”, because multiple copies of an identical molecule of DNA are produced. Another way of producing many identical copies of a particular (often short, for example, 100-3,000 base pairs) DNA fragment is the polymerase chain reaction. This method is rapid and avoids the need for cloning DNA into a vector.

Reporter gene

In molecular biology, a reporter gene is a gene that researchers attach to a regulatory sequence of another gene of interest in cell culture, animals or plants. Certain genes are chosen as reporters because the characteristics they confer on organisms expressing them are easily identified and measured, or because they are selectable markers. Reporter genes are often used as an indication of whether a certain gene has been taken up by or expressed in the cell or organism population.

To introduce a reporter gene into an organism, scientists place the reporter gene and the gene of interest in the same DNA construct to be inserted into the cell or organism. For bacteria or eukaryotic cells in culture, this is usually in the form of a circular DNA molecule called a plasmid. It is important to use a reporter

gene that is not natively expressed in the cell or organism under study, since the expression of the reporter is being used as a marker for successful uptake of the gene of interest. Commonly used reporter genes that induce visually identifiable characteristics usually involve fluorescent and luminescent proteins; examples include the gene that encodes jellyfish green fluorescent protein (GFP), which causes cells that express it to glow green under blue light, the enzyme luciferase, which catalyzes a reaction with luciferin to produce light, and the red fluorescent protein from the gene dsRed. Another common reporter in bacteria is the Lac Z gene, which encodes the protein beta-galactosidase. This enzyme causes bacteria expressing the gene to appear blue when grown on a medium that contains the substrate analog X-gal. An example of a selectable-marker reporter in bacteria is the chloramphenicol acetyltransferase (CAT) gene, which confers resistance to the antibiotic chloramphenicol.

Reporter genes can also be used to assay for the expression of the gene of interest, which may produce a protein that has little obvious or immediate effect on the cell culture or organism. In these cases the reporter is directly attached to the gene of interest to create a gene fusion. The two genes are under the same promoter elements and are transcribed into a single messenger RNA molecule. The mRNA is then translated into protein. In these cases it is important that both proteins be able to properly fold into their active conformations and interact with their substrates despite being fused. In building the DNA construct, a segment of DNA coding for a flexible polypeptide linker region is usually included so that the reporter and the gene product will only minimally interfere with one another.^[23]

Gene expression and purification in Practice

Protein expression is crucial in biochemistry as it provides the substrate or enzyme required for further analysis. Before large scale protein expression, a small scale expression checked is usually first applied. BL21 Competent E. coli is a commonly used protein expression competent cells. It contains resistant to certain

antibiotics, such as Kanamycin; it can undergo modification to express proteins of interest.

In an expression check, desired genome is inoculated and expressed overnight in 5ml appropriate media with corresponding antibiotics. After overnight expression, the media containing desired protein is spun down. After supernatant is removed, the pellet is suspended into appropriate lysis buffer and sonicated. After sonication, the sample is spun down. The soluble fraction and insoluble fractions are taken for gel analysis on an SDS page. The approximate size of the desired protein must be known and calculated ahead. If desire band shows up in the SDS page, large scale protein expression can then move on.

In a large scale protein expression, three liter cultural flasks are commonly used for inoculation and induction. At the beginning, starter culture of gene of interests needs to be prepared by inoculating already modified protein expression competent cell in 5ml to 25ml of autoclaved media. Media commonly used are LB, TB, etc. Starter culture needs to be incubated in appropriate temperature, such as 37 Celsius, along with well shaking overnight. On the same day, litters of media can be prepared. In the case of LB media, 25 grams of LB is required per liter of deionized water. The culture flasks are taped with aluminum foil and autoclaved. Before inoculation, the LB needs to be remained covered with aluminum foil to stay sterile. On the day of inoculation, the culture media needs to be cool to at least room temperature. Appropriate antibiotics needs to be added into the media with well shaking. Inoculation is done via adding 5-10ml



Affinity Chromatography

of starter culture into each liter of culture. The media is then placed in a 37 degree shaker. It is important to keep track of the optical density of the inoculate culture. Desire optical density is 0.6. In the case of E. coli, such optical density is obtained after about 3 hours of inoculation. E. coli duplicate its amount every 20 minutes; however, in the presence of antibiotics, such duplication period might take longer. But a 3-hour inoculation period is safe. After 3 hours, the optical density of media should be carefully monitored. If the OD is too low, induction might not be sufficient; while if the OD is too high, we might obtain undesired protein, so an OD of 0.6 is desired. After the OD reaches 0.6, the media needs to be chilled on ice before induction. IPTG is commonly used to induce BL21 competent cell. 1mM of IPTG is sufficient for induction. The induction temperature might be different from inoculation temperature. Induction takes place over night.

On the second day, media is spun down into pellets. The pellets needs to be lysed via either a French or a microfluidizer depending on the amount of pellet available. French Press is usually good for a 2L culture lysis, while microfluidizer is a better choice for anything go beyond. The actual choice also depends on what is available and how soluble the pellets are in lysis buffer.

After the pellets are suspended in lysis buffer, lysozyme, DNase, and RNase are added, and inculated for at least 10 minutes. If inadequate amount of any of these are added, the pellets will appear sticky during lysis and lysis might be incomplete. After lysis, the sample is spun down to obtain the soluble fraction that contain our desired protein. It is crucial to keep the whole lysing process cold as some protein might precipitate in room temperature, or machine got heated up will cause a loss of protein. With the soluble fraction that contain our protein of interest, further purification can be done. Such as salting out, ion exchange, and affinity chromatography. Further purification might involve FPLC.

The appropriate lysis buffer is crucial in protein expression. Different plasmids express differently in different media.

Temperature, and pH are also important factors to take into account.

References

1. ↑ Gene expression
2. ↑ Transcription (genetics)
3. ↑ Beadle GW, Tatum EL (15 November 1941). "Genetic Control of Biochemical Reactions in Neurospora". *PNAS* 27 (11): 499–506.
4. ↑ One gene-one enzyme hypothesis
5. ↑ Operon
6. ↑ <http://en.wikipedia.org/w/index.php?title=Riboswitch&oldid=420979855>
7. ↑ Lac operon
8. ↑ http://en.wikipedia.org/w/index.php?title=Lac_repressor&oldid=426589639
9. ↑ http://en.wikipedia.org/w/index.php?title=Trp_operon&oldid=428240917
10. ↑ Gene expression
11. ↑ Plasmid
12. ↑ Polymerase chain reaction
13. ↑ Polymerase chain reaction
14. ↑ Pingoud A, Jeltsch A (September 2001). "Structure and function of type II restriction endonucleases". *Nucleic Acids Res.* 29 (18): 3705–27. doi:10.1093/nar/29.18.3705. PMID 11557805.
15. ↑ Molecular Biology: Understanding the Genetic Revolution, by David P. Clark. Elsevier Academic Press, 2005. ISBN 0-12-175551-7.
16. ↑ Goodsell DS (2002). "The molecular perspective: restriction endonucleases". *Stem Cells* 20 (2): 190–1. doi:10.1634/stemcells.20-2-190. PMID 11897876.
17. ↑ Roberts RJ (January 1980). "Restriction and modification

- enzymes and their recognition sequences". Nucleic Acids Res. **8** (1): r63–r80. doi:10.1093/nar/8.1.197-d. PMID 6243774.
- 18. ↑ R.J Roberts, 1988, Nucl Acids Res. 16(suppl):271 From p.213 Molecular Cell Biology 4th Edition by Lodish, Berk, Zipursky, Matsudaira, Baltimore and Darnell.
 - 19. ↑ "Stu I from Streptomyces tubercidicus". Sigma-Aldrich. [http://www.sigmaaldrich.com/catalog/search/
ProductDetail/SIGMA/R8013](http://www.sigmaaldrich.com/catalog/search/ProductDetail/SIGMA/R8013). Retrieved 2008-06-07.
 - 20. ↑ Shimotsu H, Takahashi H, Saito H (November 1980). "A new site-specific endonuclease StuI from Streptomyces tubercidicus". Gene **11** (3-4): 219–25. doi:10.1016/0378-1119(80)90062-1. PMID 6260571.
 - 21. ↑ Copeland NG, Jenkins NA, Court DL (October 2001). "Recombineering: a powerful new tool for mouse functional genomics". Nat. Rev. Genet. **2** (10): 769–79. doi:10.1038/35093556. PMID 11584293.
 - 22. ↑ Lu JP, Beatty LK, Pintus JH. (2008). "Dual expression recombinase based (DERB) single vector system for high throughput screening and verification of protein interactions in living cells.". *Nature Precedings*.
 - 23. ↑ [http://en.wikipedia.org/w/
index.php?title=Reporter_gene&oldid=422027341](http://en.wikipedia.org/w/index.php?title=Reporter_gene&oldid=422027341)

I8.

Like other biological macromolecules such as polysaccharides and nucleic acids, proteins are essential parts of living organisms and participate in virtually every process within cells. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signaling, immune responses, cell adhesion, and the cell cycle. Proteins are also necessary in animals' diets, since animals cannot synthesize all the amino acids they need and must obtain essential amino acids from food. Through the process of digestion, animals break down ingested protein into free amino acids that are then used in metabolism. Proteins were first described by the Dutch chemist Gerhardus Johannes Mulder and named by the Swedish chemist Jöns Jakob Berzelius in 1838.^[1]

Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. The genetic code is a set of three-nucleotide sets called codons and each three-nucleotide combination designates an amino acid, for example AUG (Adenine-Uracil-Guanine) is the code for methionine. Because DNA contains four nucleotides, the total number of possible codons is 64; hence, there is some redundancy in the genetic code, with some amino acids specified by more than one codon. Genes encoded in DNA are first transcribed into pre-messenger RNA (mRNA) by proteins such as RNA polymerase. Most organisms then process the pre-mRNA (also known as a primary transcript) using various forms of post-transcriptional modification to form the mature mRNA, which is then used as a template for protein synthesis by the ribosome. In prokaryotes the mRNA may either be used as soon as it is produced, or be bound by a ribosome

after having moved away from the nucleoid. In contrast, eukaryotes make mRNA in the cell nucleus and then translocate it across the nuclear membrane into the cytoplasm, where protein synthesis then takes place. The rate of protein synthesis is higher in prokaryotes than eukaryotes and can reach up to 20 amino acids per second.

The mRNA is loaded onto the ribosome and is read three nucleotides at a time by matching each codon to its base pairing anticodon present on a transfer RNA (t-RNA) molecule, which carries the amino acid corresponding to the codon it recognizes. The enzyme aminoacyl tRNA synthetase “charges” the tRNA molecules with the correct amino acids. The growing polypeptide is often termed the nascent chain. Proteins are always biosynthesized from N-terminus to C-terminus.

The size of a synthesized protein can be measured by the number of amino acids it contains and by its total molecular mass, which is normally reported in units of daltons (synonymous with atomic mass units), or the derivative unit **Kilodalton (kDa)**. **Yeast proteins are on average 466 amino acids long and 53 kDa in mass**. The largest known proteins are the **titins**, a component of the muscle sarcomere, with a molecular mass of almost 3,000 kDa and a total length of almost 27,000 amino acids.

Contents

- 1 Translation: The Synthesis of protein
- 2 tRNA (Transfer RNA)
- 3 Ribosome
 - 3.1 Structure of the Ribosome
 - 3.2 Kozak consensus sequence
- 4 Synthesis of protein in Prokaryotes

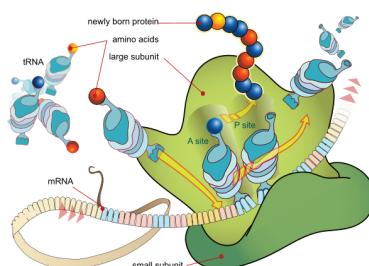


Diagram showing the translation of mRNA and the synthesis of proteins by a ribosome.

- 4.1 Initiation
- 5 Elongation
 - 5.1 Termination
 - 5.2 Recycling
 - 5.3 Polysomes
- 6 Synthesis of protein in Eukaryotes
 - 6.1 Initiation
 - 6.1.1 The cap-independent initiation
 - 6.1.2 Cap-dependent initiation
 - 6.2 Elongation
 - 6.3 Termination
- 7 Structure of protein
 - 7.1 Primary structure of protein
 - 7.2 Secondary structure of protein
 - 7.3 Tertiary structure of protein
 - 7.4 Quaternary structure of protein
- 8 Drugs which inhibit protein synthesis
- 9 Posttranslational modification of protein
- 10 Identification of protein in laboratory
 - 10.1 SDS-Gel electrophoresis
 - 10.2 Setting of transfer in semi dry assembly
- 11 Facts to be remembered
- 12 Question time
- 13 References

Translation: The Synthesis of protein

The synthesis of proteins is known as translation. **Translation generally occurs in the cytoplasm, where the ribosomes are located.** Ribosomes are made of a small and large subunit that surround the mRNA. In translation, messenger RNA (mRNA) is decoded to produce a specific polypeptide according to the rules

specified by the trinucleotide genetic code. This uses an mRNA sequence as a template to guide the synthesis of a chain of amino acids that form a protein. Translation proceeds in four phases: activation, initiation, elongation, and termination (all describing the growth of the amino acid chain, or polypeptide that is the product of translation).

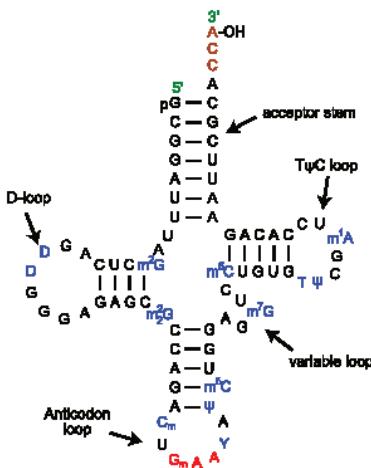
In activation, the correct amino acid (AA) is joined to the correct transfer RNA (tRNA). While this is not technically a step in translation, it is required for translation to proceed. The AA is joined by its carboxyl group to the 3' OH of the tRNA by an ester bond. When the tRNA has an amino acid linked to it, it is termed “charged”. Initiation involves the small subunit of the ribosome binding to 5' end of mRNA with the help of initiation factors (IF), other proteins that assist the process. Elongation occurs when the next aminoacyl-tRNA (charged tRNA) in line binds to the ribosome along with GTP and an elongation factor. Termination of the polypeptide happens when the A site of the ribosome faces a **stop codon (UAA, UAG, or UGA)**. When this happens, no tRNA can recognize it, but releasing factor can recognize nonsense codons and causes the release of the polypeptide chain. The capacity of disabling or inhibiting translation in protein biosynthesis is used by antibiotics such as: anisomycin, cycloheximide, chloramphenicol, tetracycline, streptomycin, erythromycin, puromycin etc.^[2] Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. The genetic code is a set of three-nucleotide sets called codons and each three-nucleotide combination designates an amino acid, for example AUG (Adenine-Uracil-Guanine) is the code for methionine. Because DNA contains four nucleotides, the total number of possible codons is 64; hence, there is some redundancy in the genetic code, with some amino acids specified by more than one codon. Genes encoded in DNA are first transcribed into pre-messenger RNA (mRNA) by proteins such as RNA polymerase. Most organisms then process the pre-mRNA (also known as a primary transcript) using various forms of post-transcriptional modification

to form the mature mRNA, which is then used as a template for protein synthesis by the ribosome. In prokaryotes the mRNA may either be used as soon as it is produced, or be bound by a ribosome after having moved away from the nucleoid. In contrast, eukaryotes make mRNA in the cell nucleus and then translocate it across the nuclear membrane into the cytoplasm, where protein synthesis then takes place. The rate of protein synthesis is higher in prokaryotes than eukaryotes and can reach up to 20 amino acids per second. We should always remember that following antibiotics inhibit the protein synthesis e.g. **anisomycin**, **cycloheximide**, **chloramphenicol**, **tetracycline**, **streptomycin**, **erythromycin**, **puromycin** etc.

The mRNA is loaded onto the ribosome and is read three nucleotides at a time by matching each codon to its base pairing anticodon present on a transfer RNA (t-RNA) molecule, which carries the amino acid corresponding to the codon it recognizes. The enzyme aminoacyl tRNA synthetase “charges” the tRNA molecules with the correct amino acids. The growing polypeptide is often termed the nascent chain. Proteins are always biosynthesized from N-terminus to C-terminus. The size of a synthesized protein can be measured by the number of amino acids it contains and by its total molecular mass, which is normally reported in units of daltons (synonymous with atomic mass units), or the derivative unit **Kilodalton (kDa)**. **Yeast proteins are on average 466 amino acids long and 53 kDa in mass.** The largest known proteins are the **titins**, a component of the muscle sarcomere, with a molecular mass of almost 3,000 kDa and a total length of almost 27,000 amino acids.^[3]

tRNA (Transfer RNA)

tRNA appears like cloverleaf structure. Its anticodon arm is a 5 base pair (bp) stem whose loop contains the anticodon while its D arm is a 4 bp stem ending in a loop that often contains dihydrouridine. T arm of tRNA is a 5 bp stem containing the sequence TΨC where Ψ is a pseudouridine. In eukaryotic cells, tRNAs are transcribed by RNA pol III as pre-tRNAs in the nucleus. tRNA is a small RNA molecule usually 73-95 nucleotides long. RNA polymerase III recognizes two internal promoter sequences (A-box B internal promoter) inside tRNA genes. The first promoter begins at nucleotide 8 of mature tRNAs and the second promoter is located 30-60 nucleotides downstream of the first promoter. The transcription terminates after a stretch of four or more thymidines.



Secondary cloverleaf structure of tRNA^{Phe} from yeast.

Pre-tRNAs undergo extensive modifications inside the nucleus. Some pre-tRNAs contain introns; in bacteria these self-splice, whereas in eukaryotes and archaea they are removed by tRNA splicing endonuclease. The 5' sequence is removed by RNase P, whereas the 3' end is removed by the tRNase Z enzyme. A notable exception is in the archaeon *Nanoarchaeum equitans* which does not possess an RNase P enzyme and has a promoter placed such that transcription starts at the 5' end of the mature tRNA. The non-templated 3' CCA tail is added by a nucleotidyl transferase. Before tRNAs are exported into the cytoplasm by Los1/Xpo-t, tRNAs are aminoacylated. The order of the processing events is not conserved.

For example in yeast, the splicing is not carried out in the nucleus but at the cytoplasmic side of mitochondrial membranes.^[4]

Ribosome

Ribosomes are the components of cells that make proteins from all amino acids. One of the central tenets of biology, often referred to as the “central dogma,” is that DNA is used to make RNA, which, in turn, is used to make protein. The DNA sequence in genes is copied into a messenger RNA (mRNA). Ribosomes then read the information in this RNA and use it to create proteins. This process is known as translation; i.e., the ribosome “translates” the genetic information from RNA into proteins. Ribosomes do this by binding to an mRNA and using it as a template for the correct sequence of amino acids in a particular protein. The amino acids are attached to transfer RNA (tRNA) molecules, which enter one part of the ribosome and bind to the messenger RNA sequence. The attached amino acids are then joined together by another part of the ribosome. The ribosome moves along the mRNA, “reading” its sequence and producing a chain of amino acids.

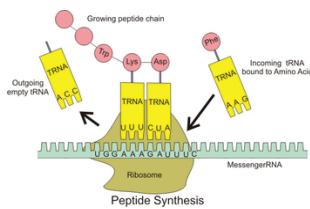
Ribosomes are made from complexes of RNAs and proteins. Ribosomes are divided into two subunits, one larger than the other. The smaller subunit binds to the mRNA, while the larger subunit binds to the tRNA and the amino acids. When a ribosome finishes reading a mRNA, these two subunits split apart. Ribosomes have been classified as ribozymes, since the ribosomal RNA seems to be most important for the peptidyl transferase activity that links amino acids together.^[5]

Ribosomes from bacteria, archaea and eukaryotes (the three domains of life on Earth), have significantly different structures and RNA sequences. These differences in structure allow some antibiotics to kill bacteria by inhibiting their ribosomes, while leaving human ribosomes unaffected. The ribosomes in the

mitochondria of eukaryotic cells resemble those in bacteria, reflecting the likely evolutionary origin of this organelle. **The word ribosome comes from ribonucleic acid and the Greek: soma (meaning body).**

Structure of the Ribosome

A svedberg (symbol S, sometimes Sv, not to be confused with Sv for the SI unit sievert as well as the non-SI sverdrup) is a non-SI physical unit used for sedimentation coefficients. It characterizes the behaviour



Ribosomes read the sequence of messenger RNAs and assemble proteins out of amino acids bound to transfer RNAs.

of a particle type in sedimentation processes, notably centrifugation. The svedberg is technically a measure of time, and is defined as exactly 10⁻¹³ seconds (100 fs). The unit is named after the Swedish chemist Theodor Svedberg (1884–1971), winner of the Nobel prize in chemistry in 1926 for his work in the chemistry of colloids and his invention of the ultracentrifuge.

Bigger particles tend to sediment faster and thus have higher svedberg values. Sedimentation coefficients are, however, not additive. Sedimentation rate does not depend only on the mass or volume of a particle, and when two particles bind together there is inevitably a loss of surface area. Thus when measured separately they will have svedberg values that may not add up to that of the bound particle. The svedberg is the most

important measure used to distinguish ribosomes, which are important in phylogenetic studies.

The ribosomal subunits of prokaryotes and eukaryotes are quite similar. The unit of measurement is the Svedberg unit, a measure of the rate of sedimentation in centrifugation rather than size and accounts for why fragment names do not add up (70S is made of 50S and 30S). **Prokaryotes have 70S ribosomes, each consisting of a small (30S) and a large (50S) subunit.** Their large subunit is composed of a 5S RNA subunit (consisting of 120 nucleotides), a 23S RNA subunit (2900 nucleotides) and 34 proteins. The 30S subunit has a 1540 nucleotide RNA subunit (16S) bound to 21 proteins. **Eukaryotes have 80S ribosomes, each consisting of a small (40S) and large (60S) subunit.** Their large subunit is composed of a 5S RNA (120 nucleotides), a 28S RNA (4700 nucleotides), a 5.8S subunit (160 nucleotides) and ~49 proteins. The 40S subunit has a 1900 nucleotide (18S) RNA and ~33 proteins.

The ribosomes found in chloroplasts and mitochondria of eukaryotes also consist of large and small subunits bound together with proteins into one 70S particle. These organelles are believed to be descendants of bacteria (see Endosymbiotic theory) and as such their ribosomes are similar to those of bacteria. The various ribosomes share a core structure, which is quite similar despite the large differences in size. Much of the RNA is highly organized into various tertiary structural motifs, for example pseudoknots that exhibit coaxial stacking. The extra RNA in the larger ribosomes is in several long continuous insertions, such that they form loops out of the core structure without disrupting or changing it. All of the catalytic activity of the ribosome is carried out by the RNA; the proteins reside on the surface and seem to stabilize the structure.

The differences between the bacterial and eukaryotic ribosomes are exploited by pharmaceutical chemists to create antibiotics that can destroy a bacterial infection without harming the cells of the infected person. Due to the differences in their structures, the

bacterial 70S ribosomes are vulnerable to these antibiotics while the eukaryotic 80S ribosomes are not. Even though mitochondria possess ribosomes similar to the bacterial ones, mitochondria are not affected by these antibiotics because they are surrounded by a double membrane that does not easily admit these antibiotics into the organelle.

Ribogenesis in Prokaryotes

There are 52 genes that encode the ribosomal proteins and they can be found in 20 operons within prokaryotic DNA. Regulation of ribosome synthesis hinges on the regulation of the rRNA itself. First, a reduction in aminoacyl-tRNA will cause the prokaryotic cell to respond by lowering transcription and translation. This occurs through a series of steps, beginning with stringent factor binding to ribosomes and catalyzing the reaction:

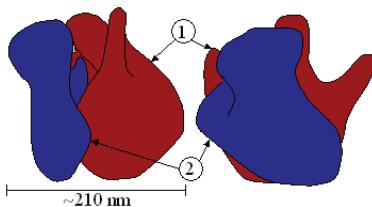
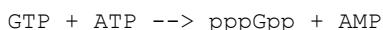


Figure : The large subunit red (1) and small subunit in blue(2)

The γ -phosphate is then removed and ppGpp will bind to and inhibit RNA polymerase. This binding causes a reduction in rRNA transcription. A reduced amount of rRNA means that ribosomal proteins (r-proteins) will be translated but will not have an rRNA to bind to. Instead, they will negatively feedback and bind to their own mRNA, repressing r-protein

synthesis. Note that r-proteins preferentially bind to its complementary rRNA if it is present, rather than mRNA. The ribosome operons also include the genes for RNA polymerase and elongation factors (used in RNA translation). Regulation of all of these genes at once illustrate the coupling between transcription and translation in prokaryotes.

Ribogenesis in Eukaryotes

Ribosomal protein synthesis in eukaryotes occurs, like most protein

synthesis, in the cytoplasm just outside the nucleus. Individual large and small units are synthesized and imported into the nucleus through nuclear pores. These pores have a diameter of 120 nm and import 560,000 ribosomal proteins per minute into the nucleus with active transport. See nuclear import for more about the movement of the ribosomal proteins into the nucleus. Ribosomal RNA (rRNA) is transcribed at the nucleolus, at a high speed, which contains all 45S rRNA genes. After transcription, the rRNA is put together with the ribosomal subunits to make a functioning ribosome.

Kozak consensus sequence

Kozak consensus sequence on an mRNA molecule is recognized by the ribosome as the translational start site, from which a protein is coded by that mRNA molecule. The ribosome requires this sequence, or a possible variation to initiate translation. The Kozak sequence is not to be confused with the ribosomal binding site (**RBS**), that being either the 5' cap of a messenger RNA or an Internal Ribosome Entry Site (IRES). In vivo, this site is often not matched exactly on different mRNAs and the amount of protein synthesized from a given mRNA is dependent on the strength of the Kozak sequence. Some nucleotides in this sequence are more important than others: the AUG is most important because it is the actual initiation codon encoding a methionine amino acid at the N-terminus of the protein. (Rarely, CTG is used as an initiation codon, encoding a leucine instead of its typical methionine.) The A nucleotide of the “AUG” is referred to as number 1. For a ‘strong’ consensus, the nucleotides at positions +4 (i.e. G in the consensus) and -3 (i.e. either A or G in the consensus) relative to the number 1 nucleotide must both match the consensus (there is no number 0 position). An ‘adequate’ consensus has only 1 of these sites, while a ‘weak’ consensus has neither. The cc at -1 and -2 are not as

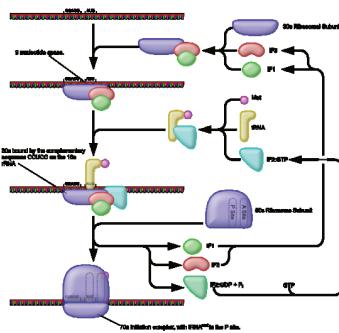
conserved, but contribute to the overall strength. There is also evidence that a G in the -6 position is important in the initiation of translation.

There are examples *in vivo* of each of these types of Kozak consensus, and they probably evolved as yet another mechanism of gene regulation. Lmx1b is an example of a gene with a weak Kozak consensus sequence. For initiation of translation from such a site, other features are required in the mRNA sequence in order for the ribosome to recognize the initiation codon.^[6]

Synthesis of protein in Prokaryotes

Initiation

Initiation of translation in prokaryotes involves the assembly of the components of the translation system which are: the two ribosomal subunits (50S & 30S subunits), the mRNA to be translated, the first (formyl) aminoacyl tRNA (the tRNA charged with the first amino acid), GTP (as a source of energy), and three initiation factors (Prokaryotic initiation factor-1 or IF1, Prokaryotic initiation factor-2 or IF2, and Prokaryotic initiation factor-3 or IF3 which help the assembly of the initiation complex.^[7] The ribosome has three active site|sites: the A site, the P site, and the E site. The A site is the point of entry for the aminoacyl tRNA (except for the first aminoacyl tRNA, fMet-tRNA_f^{Met}, which enters at the P site). The P



The process of initiation of translation in prokaryotes.

site is where the peptidyl tRNA is formed in the ribosome. And the E site which is the exit site of the now uncharged tRNA after it gives its amino acid to the growing peptide chain.

Elongation

Elongation of the polypeptide chain involves addition of amino acids to the carboxyl end of the growing chain. The growing protein exits the ribosome through the polypeptide exit tunnel in the large subunit.^[8] In prokaryotes, three elongation factors are required for translation: EF-Tu, EF-Ts, and EF-G.

EF-Tu (elongation factor thermo unstable) mediates the entry of the aminoacyl tRNA into a free site of the ribosome.

EF-Ts serves as the guanine nucleotide exchange factor for EF-Tu, catalyzing the release of GDP from EF-Tu.

EF-G catalyzes the translocation of the tRNA and mRNA down the ribosome at the end of each round of polypeptide elongation.

Elongation starts when the fmet-tRNA enters the P site, causing a conformational change which opens the A site for the new aminoacyl-tRNA to bind. This binding is facilitated by elongation factor-Tu (EF-Tu), a small GTPase. Now the P site contains the beginning of the peptide chain of the protein to be encoded and the A site has the next amino acid to be added to the peptide chain. The growing polypeptide connected to the tRNA in the P site is detached from the tRNA in the P site and a peptide bond is formed between the last amino acids of the polypeptide and the amino acid still attached to the tRNA in the A site. This process, known as *peptide bond formation*, is catalyzed by a ribozyme (the 23S ribosomal RNA in the 50S ribosomal subunit). Now, the A site has the newly formed peptide, while the P site has an uncharged tRNA (tRNA with no amino acids). In the final stage of elongation, *translocation*, the ribosome moves 3 nucleotides towards the 3'end of mRNA. Since tRNAs are linked to mRNA by codon-anticodon base-pairing, tRNAs

move relative to the ribosome taking the nascent polypeptide from the A site to the P site and moving the uncharged tRNA to the E exit site. This process is catalyzed by elongation factor G (EF-G).

The ribosome continues to translate the remaining codons on the mRNA as more aminoacyl-tRNA bind to the A site, until the ribosome reaches a stop codon on mRNA(UAA, UGA, or UAG).^[9]

EF-Tu

EF-Tu (elongation factor thermo unstable) is one of the prokaryotic elongation factors. The prokaryotic factor EF-Tu mediates the entry of the aminoacyl tRNA into a free site of the ribosome. EF-Tu functions by binding an aminoacylated, or charged, tRNA molecule in the cytoplasm. This complex transiently enters the ribosome, with the tRNA anticodon domain associating with the mRNA codon in the ribosomal A site. If the codon-anticodon pairing is correct, EF-Tu hydrolyzes guanosine triphosphate (GTP) into guanosine diphosphate (GDP) and inorganic phosphate, and changes in conformation to dissociate from the tRNA molecule. The aminoacyl tRNA then fully enters the A site, where its amino acid is brought near the P-site polypeptide and the ribosome catalyzes the covalent transfer of the amino acid onto the polypeptide. EF-Tu contributes to translational accuracy in three ways. It delays GTP hydrolysis if the tRNA in the ribosome's A site does not match the mRNA codon, thus preferentially increasing the likelihood for the incorrect tRNA to leave the ribosome. It also adds a second delay (regardless of tRNA matching) after freeing itself from tRNA, before the aminoacyl tRNA fully enters the A site. This delay period is a second opportunity for incorrectly-paired tRNA (and their bound amino acids) to move out of the A site before the incorrect amino acid is irreversibly added to the polypeptide chain. A third mechanism is the less well understood function of EF-Tu to crudely check amino acid-tRNA associations, and reject complexes where the amino acid is not bound to the correct tRNA coding for it.

EF-Ts

EF-Ts (elongation factor thermo stable) is one of the prokaryotic elongation factors. EF-Ts serves as the guanine nucleotide

exchange factor for EF-Tu (elongation factor thermo unstable), catalyzing the release of guanosine diphosphate from EF-Tu. This enables EF-Tu to bind to a new guanosine triphosphate molecule, release EF-Ts, and go on to catalyze another aminoacyl tRNA addition.

EF-G

The factor EF-G catalyzes the translocation of the tRNA and mRNA down the ribosome at the end of each round of polypeptide elongation. Homologous to EF-Tu + tRNA, EF-G also binds to the ribosome in its GTP-bound state. When it associates with the A site, EF-G causes the tRNA previously occupying that site to occupy an intermediate A/P position (bound to the A site of the small ribosomal subunit and to the P site of the large subunit), and the tRNA in the P site is shifted to a P/E hybrid state. EF-G hydrolysis of GTP causes a conformation change that forces the A/P tRNA to fully occupy the P site, the P/E tRNA to fully occupy the E site (and exit the ribosome complex), and the mRNA to shift three nucleotides down relative to the ribosome due to its association with these tRNA molecules. The GDP-bound EF-G molecule then dissociates from the complex, leaving another free A-site where the elongation cycle can start again. Apart from its role in translocation, EF-G, working together with Ribosome Recycling Factor promotes ribosome recycling in a GTP-dependent manner

Termination

Termination occurs when one of the three termination codons moves into the A site. These codons are not recognized by any tRNAs. Instead, they are recognized by proteins called release factors, namely RF1 (recognizing the UAA and UAG stop codons) or

RF2 (recognizing the UAA and UGA stop codons). These factors trigger the hydrolysis of the ester bond in peptidyl-tRNA and the release of the newly synthesized protein from the ribosome. A third release factor RF-3 catalyzes the release of RF-1 and RF-2 at the end of the termination process.

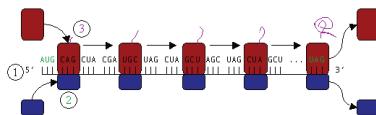


Figure : Translation of mRNA (1) by a ribosome (2)(shown as small (in blue color) and large(in red color) subunits) into a polypeptide chain (3). The ribosome begins at the start codon of mRNA (AUG) and ends at the stop codon (UAG).

Recycling

The post-termination complex formed by the end of the termination step consists of mRNA with the termination codon at the A-site, an uncharged tRNA in the P site, and the intact 70S ribosome. Ribosome recycling step is responsible for the disassembly of the post-termination ribosomal complex.^[10] Once the nascent protein is released in termination, Ribosome Recycling Factor and Elongation Factor G (EF-G) function to release mRNA and tRNAs from ribosomes and dissociate the 70S ribosome into the 30S and 50S subunits. IF3 then replaces the deacylated tRNA releasing the mRNA. All translational components are now free for additional rounds of translation.

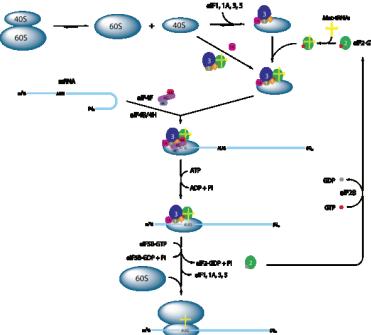
Polysomes

Translation is carried out by more than one ribosome simultaneously. Because of the relatively large size of ribosomes, they can only attach to sites on mRNA 35 nucleotides apart. The complex of one mRNA and a number of ribosomes is called a polysome or polyribosome.

Synthesis of protein in Eukaryotes

Initiation

Eukaryotic initiation factors (eIF) are proteins involved in the initiation phase of eukaryotic translation. They function in forming a complex with the 40S ribosomal subunit and Met-tRNA_i called the 43S preinitiation complex (PIC), recognizing the 5' cap structure of mRNA and recruiting the 43S PIC to mRNA, promoting ribosomal scanning of mRNA and regulating recognition of the AUG initiation codon, and joining of the 60S ribosomal subunit to create the 80S ribosome. There exist many more eukaryotic initiation factors than prokaryotic initiation factors due to greater biological complexity of eukaryotic cells. The protein RLI is known to have an essential, probably catalytic role in the formation of initiation complexes as well.



The process of initiation of translation in eukaryotes with eIF2 in light green. Other factors are shown too.

EIF4 (eIF4F)

The eIF4 initiation factors include eIF4A, eIF4B, eIF4E, and eIF4G. eIF4F is often used to refer to the complex of eIF4A, eIF4E, and eIF4G. eIF4G is a scaffolding protein that interacts with eIF3 (see below), as well as the other members of the eIF4F complex. eIF4E recognizes and binds to the 5' cap structure of mRNA, while eIF4G binds to Poly(A)-binding protein which binds the poly(A) tail, circularizing and activating the bound mRNA. eIF4A – a DEAD box RNA helicase – is important for resolving mRNA secondary structures. eIF4B contains two RNA-binding domains – one non-specifically interacts with mRNA, whereas the second specifically binds the 18S portion of the small ribosomal subunit. It acts as an anchor, as well as a critical co-factor for eIF4A. It is a substrate of S6K, and, when phosphorylated, it promotes the formation of the pre-initiation complex. In vertebrates, eIF4H is an additional initiation factor with similar function to eIF4B.^[11]

EIF1 & eIF3

eIF1, eIF1A, and eIF3, all bind to the ribosome subunit-mRNA complex. They have been implicated in preventing the large ribosomal subunit from binding the small subunit before it is ready to commence elongation. In mammals, eIF3 is the largest scaffolding initiation factor, made up of 13 subunits (a-m). It is roughly ~750 kDa and it controls the assembly of 40S ribosomal subunit on mRNA that have a 5' cap or an IRES. eIF3 uses the eIF4F complex or IRES (Internal Ribosomal Entry Site) from viruses to position the mRNA strand near the exit site of the 40S ribosome subunit, thus promoting the assembly of the pre-initiation complex. In many cancers, eIF3 is overexpressed. Under serum-deprived conditions (inactive state), eIF3 is bound to S6K1. On stimulation either by mitogens, growth factors, or drugs, mTOR/Raptor complex gets activated and, in turn, binds and phosphorylates S6K1 on T389 (linker region), causing a conformational change that causes the kinase S6K1 to dissociate from eIF3. The T389 phosphorylated S6K1 is then further phosphorylated by PDK1 on T229. This second

phosphorylation fully activates the S6K1 kinase, which can then phosphorylate eIF4B, S6 and other protein targets.^[12]

eIF2

eIF2 is a GTP-binding protein responsible for bringing the initiator tRNA to the P-site of the pre-initiation complex. It has specificity for the methionine-charged initiator tRNA, which is distinct from other methionine-charged tRNAs specific for elongation of the polypeptide chain. Once it has placed the initiator tRNA on the AUG start codon in the P-site, it hydrolyzes GTP into GDP, and dissociates. This hydrolysis, also signals for the dissociation of eIF3, eIF1, and eIF1A, and allows the large subunit to bind. This signals the beginning of elongation. eIF2 has three subunits, eIF2- α , β , and γ . The former is of particular importance for cells that may need to turn off protein synthesis globally. When phosphorylated, it sequesters eIF2B (not to be confused with beta), a GEF. Without this GEF, GDP cannot be exchanged for GTP, and translation is repressed. eIF2 α -induced translation repression occurs in reticulocytes when starved for iron. In addition, protein kinase R (PKR) phosphorylates eIF2 α when dsRNA is detected in many multicellular organisms, leading to cell death.^[13]

eIF5 & eIF5B

eIF5A is a GTPase-activating protein, which helps the large ribosomal subunit associate with the small subunit. It is required for GTP-hydrolysis by eIF2 and contains the unusual amino acid hypusine. eIF5B is a GTPase, and is involved in assembly of the full ribosome (which requires GTP hydrolysis).

eIF6

eIF6 performs the same inhibition of ribosome assembly as eIF3, but binds with the large subunit.

The process of initiation of translation in eukaryotes depend up on mRNA capping.

The cap-independent initiation

The best studied example of the cap-independent mode of translation initiation in eukaryotes is the Internal Ribosome Entry Site (IRES) approach. What differentiates cap-independent translation from cap-dependent translation is that cap-independent translation does not require the ribosome to start scanning from the 5' end of the mRNA cap until the start codon. The ribosome can be trafficked to the start site by ITAFs (IRES trans-acting factors) bypassing the need to scan from the 5' end of the untranslated region of the mRNA. This method of translation has been recently discovered, and has found to be important in conditions that require the translation of specific mRNAs, despite cellular stress or the inability to translate most mRNAs. Examples include factors responding to apoptosis, stress-induced responses.

Cap-dependent initiation

Initiation of translation usually involves the interaction of certain key proteins with a special tag bound to the 5'-end of an mRNA molecule, the 5' cap. The protein factors bind the small ribosomal subunit (also referred to as the 40S subunit), and these initiation factors hold the mRNA in place. The eukaryotic Initiation Factor 3 (eIF3) is associated with the small ribosomal subunit, and plays a role in keeping the large ribosomal subunit from prematurely binding. eIF3 also interacts with the eIF4F complex which consists of three other initiation factors: eIF4A, eIF4E and eIF4G. eIF4G is a scaffolding protein which directly associates with both eIF3 and the other two components. eIF4E is the cap-binding protein. It is the rate-limiting step of cap-dependent initiation, and is often cleaved from the complex by some viral proteases to limit the cell's ability to translate its own transcripts. This is a method of hijacking the host machinery in favor of the viral (cap-independent) messages.

eIF4A is an ATP-dependent RNA helicase, which aids the ribosome in resolving certain secondary structures formed by the mRNA transcript. There is another protein associated with the eIF4F complex called the Poly(A)-binding protein (PABP), which binds the poly-A tail of most eukaryotic mRNA molecules. This protein has been implicated in playing a role in circularization of the mRNA during translation. This pre-initiation complex (43S subunit, or the 40S and mRNA) accompanied by the protein factors move along the mRNA chain towards its 3'-end, scanning for the 'start' codon (typically AUG) on the mRNA, which indicates where the mRNA will begin coding for the protein. In eukaryotes and archaea, the amino acid encoded by the start codon is methionine. The initiator tRNA charged with Met forms part of the ribosomal complex and thus all proteins start with this amino acid (unless it is cleaved away by a protease in subsequent modifications). The Met-charged initiator tRNA is brought to the P-site of the small ribosomal subunit by eukaryotic Initiation Factor 2 (eIF2). It hydrolyzes GTP, and signals for the dissociation of several factors from the small ribosomal subunit which results in the association of the large subunit (or the 60S subunit). The complete ribosome (80S) then commences translation elongation, during which the sequence between the 'start' and 'stop' codons is translated from mRNA into an amino acid sequence—thus a protein is synthesized.

Regulation of protein synthesis is dependent on phosphorylation of initiation factor eIF2 which is a part of the met-tRNA_i complex. When large numbers of eIF2 are phosphorylated, protein synthesis is inhibited. This would occur if there is amino acid starvation or there has been a virus infection. However naturally a small percentage is of this initiation factor is phosphorylated. Another regulator is 4EBP which binds to the initiation factor eIF4E found on the 5' cap on mRNA stopping protein synthesis. To oppose the effects of the 4EBP growth factors phosphorylate 4EBP reducing its affinity for eIF4E and permitting protein synthesis.^[14]

Elongation

Eukaryotic elongation factors are very similar to those in prokaryotes. Elongation in eukaryotes is carried out with two elongation factors: eEF-1 and eEF-2. The first is eEF-1, and has two subunits, α and $\beta\gamma$. α acts as counterpart to prokaryotic EF-Tu, mediating the entry of the aminoacyl tRNA into a free site of the ribosome. $\beta\gamma$ acts as counterpart to prokaryotic EF-Ts, serving as the guanine nucleotide exchange factor for α , catalyzing the release of GDP from α . The second elongation factor is eEF-2, the counterpart to prokaryotic EF-G, catalyzing the translocation of the tRNA and mRNA down the ribosome at the end of each round of polypeptide elongation.

At the end of the initiation step, the mRNA is positioned so that the next codon can be translated during the elongation stage of protein synthesis. The initiator tRNA occupies the P site in the ribosome, and the A site is ready to receive an aminoacyl-tRNA. During chain elongation, each additional amino acid is added to the nascent polypeptide chain in a three-step microcycle. The steps in this microcycle are (1) positioning the correct aminoacyl-tRNA in the A site of the ribosome, (2) forming the peptide bond and (3) shifting the mRNA by one codon relative to the ribosome. The translation machinery works relatively slowly compared to the enzyme systems that catalyze DNA replication. Proteins in prokaryotes are synthesized at a rate of only 18 amino acid residues per second, whereas bacterial replisomes synthesize DNA at a rate of 1000 nucleotides per second. This difference in rate reflects, in part, the difference between polymerizing four types of nucleotides to make nucleic acids and polymerizing 20 types of amino acids to make proteins. Testing and rejecting incorrect aminoacyl-tRNA molecules takes time and slows protein synthesis. The rate of transcription in prokaryotes is approximately 55 nucleotides per second, which corresponds to about 18 codons per second, or the same rate at which the mRNA is translated. In

bacteria, translation initiation occurs as soon as the 5' end of an mRNA is synthesized, and translation and transcription are coupled. This tight coupling is not possible in eukaryotes because transcription and translation are carried out in separate compartments of the cell (the nucleus and cytoplasm). Eukaryotic mRNA precursors must be processed in the nucleus (e.g. capping, polyadenylation, splicing) before they are exported to the cytoplasm for translation.^[15]

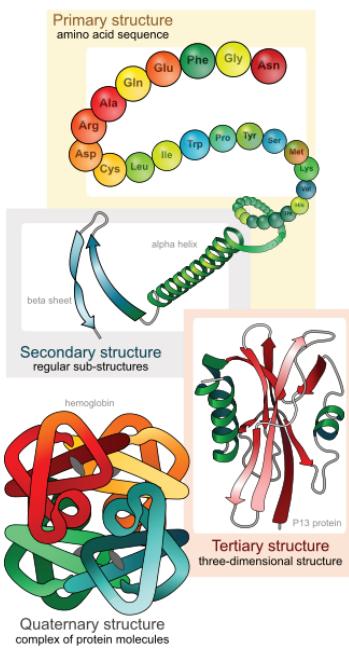
Termination

Termination of elongation is dependent on eukaryotic release factors. The process of termination is similar to that of prokaryotic termination.

Structure of protein

Primary structure of protein

The primary structure refers to the sequence of the different amino acids of the peptide or protein. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity. Counting of residues always starts at the N-terminal end (NH_2 -group), which is the end where the amino group is not involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein.



Protein structure, from **primary** to **quaternary structure**.

A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as Edman degradation or tandem mass spectrometry. Often however, it is read directly from the sequence of the gene using the genetic code. Post-translational

modifications such as disulfide formation, phosphorylations and glycosylations are usually also considered a part of the primary structure, and cannot be read from the gene.^[16]

Secondary structure of protein

In molecular biology and structural biology, secondary structure is the general three-dimensional form of local segments of biopolymers such as proteins and nucleic acids (DNA/RNA). It does not, however, describe specific atomic positions in three-dimensional space, which are considered to be tertiary structure.

Secondary structure can be formally defined by the hydrogen bonds of the biopolymer, as observed in an atomic-resolution structure. In proteins, the secondary structure is defined by the patterns of hydrogen bonds between backbone amide and carboxyl groups. In nucleic acids, the secondary structure is defined by the hydrogen bonding between the nitrogenous bases. The hydrogen bonding patterns may be significantly distorted, which makes an automatic determination of secondary structure difficult. The secondary structure may be also defined based on the regular pattern of backbone dihedral angles in a particular region of the Ramachandran plot; thus, a segment of residues with such dihedral angles may be called a helix, regardless of whether it has the correct hydrogen bonds. The secondary structure may be also provided by crystallographers in the corresponding PDB file.

The rough secondary-structure content of a biopolymer (e.g., “this protein is 40% α -helix and 20% β -sheet.”) can often be estimated spectroscopically. For proteins, a common method is far-ultraviolet (far-UV, 170–250 nm) circular dichroism. A pronounced double minimum at 208 and 222 nm indicate **α -helical structure**, whereas a single minimum at 204 nm or 217 nm reflects random-coil or **β -sheet** structure, respectively. A less common method is infrared spectroscopy, which detects differences in the bond

oscillations of amide groups due to hydrogen-bonding. Finally, secondary-structure contents may be estimated accurately using the chemical shifts of an unassigned NMR spectrum. Secondary structure was introduced by **Kaj Ulrik Linderstrøm-Lang** at Stanford in 1952.^[17]

Structural features of the three major forms of protein helices

Geometry attribute	α -helix	β_{10} helix	π -helix
Residues per turn	3.6	3.0	4.4
Translation per residue	1.5 Å	2.0 Å	1.1 Å
Radius of helix	2.3 Å	1.9 Å	2.8 Å
Pitch	5.4 Å	6.0 Å	4.8 Å

Tertiary structure of protein

Tertiary structure refers to three-dimensional structure of a single protein molecule. The alpha-helices and beta-sheets are folded into a compact globule. The folding is driven by the non-specific hydrophobic interactions (the burial of hydrophobic residues from water), but the structure is stable only when the parts of a protein domain are locked into place by specific tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds. The disulfide bonds are extremely rare in cytosolic proteins, since the cytosol is generally a reducing environment.

Quaternary structure of protein

Quaternary structure is a larger assembly of several protein molecules or polypeptide chains, usually called subunits in this

context. The quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure. Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers. Specifically it would be called a dimer if it contains two subunits, a trimer if it contains three subunits, and a tetramer if it contains four subunits. The subunits are frequently related to one another by symmetry operations, such as a 2-fold axis in a dimer. Multimers made up of identical subunits are referred to with a prefix of “homo-” (e.g. a homotetramer) and those made up of different subunits are referred to with a prefix of “hetero-” (e.g. a heterotetramer, such as the two alpha and two beta chains of **hemoglobin**). Many proteins do not have the quaternary structure and function as monomers.

Drugs which inhibit protein synthesis

In general, protein synthesis inhibitors work at different stages of prokaryotic mRNA translation into proteins, like initiation, elongation (including aminoacyl tRNA entry, proofreading, peptidyl transfer, and ribosomal translocation) and termination:

Earlier stages

Rifampicin inhibits prokaryotic DNA transcription into mRNA by inhibiting DNA-dependent RNA polymerase by binding its beta-subunit.

Initiation

Linezolid acts at the initiation stage, probably by preventing the formation of the initiation complex, although the mechanism is not fully understood.

Aminoacyl tRNA entry

Tetracyclines and Tigecycline(a glycylcycline related to tetracyclines) block the A site on the ribosome, preventing the binding of aminoacyl tRNAs.

Proofreading

Aminoglycosides, among other potential mechanisms of action, interfere with the proofreading process, causing increased rate of error in synthesis with premature termination.

Peptidyl transfer

Chloramphenicol blocks the peptidyl transfer step of elongation on the 50S ribosomal subunit in both bacteria and mitochondria. Macrolides (as well as inhibiting ribosomal translocation and other potential mechanisms) bind to the 50s ribosomal subunits, inhibiting peptidyl transfer. Quinupristin/dalfopristin act synergistically, with dalfopristin, enhancing the binding of quinupristin, as well as inhibiting peptidyl transfer. Quinupristin binds to binds to a nearby site on the 50S ribosomal subunit and prevents elongation of the polypeptide, as well as causing incomplete chains to be released.

Ribosomal translocation

Clindamycin, among other potential mechanisms. Aminoglycosides and macrolides among other potential mechanisms of action, have evidence of inhibition of ribosomal translocation. Fusidic acid prevents the turnover of elongation factor G (EF-G) from the ribosome.

Termination

Puromycin has a structure similar to that of the tyrosinyl aminoacyl-tRNA. Thus, it binds to the ribosomal A site and participates in peptide bond formation, producing peptidyl-puromycin. However, it does not engage in translocation and quickly dissociates from the ribosome, causing a premature termination of polypeptide synthesis. Macrolides and clindamycin (both also having other potential mechanisms) cause premature dissociation of the peptidyl-tRNA from the ribosome. Streptogramins also cause premature release of the peptide chain.

Protein synthesis inhibitors of unspecified mechanism

Retapamulin

Binding site

The following antibiotics bind to the 30S subunit of the ribosome:

Aminoglycosides

Tetracyclines

The following antibiotics bind to the 50S ribosomal subunit:

Chloramphenicol

Erythromycin

Clindamycin

Linezolid

Telithromycin

Streptogramins

Retapamulin

Posttranslational modification of protein

Posttranslational modification (PTM) is the chemical modification of a protein after its translation. It is one of the later steps in protein biosynthesis for many proteins.

PTMs involving addition of functional groups

PTMs involving addition by an enzyme *in vivo* acylation, e.g. O-acylation (esters), N-acylation (amides), S-acylation (thioesters) acetylation, the addition of an acetyl group, either at the N-terminus of the protein or at lysine residues. See also histone acetylation. The reverse is called deacetylation. formylation lipoylation, attachment of a lipoate (C8) functional group myristoylation, attachment of myristate, a C14 saturated acid palmitoylation, attachment of palmitate, a C16 saturated acid alkylation, the addition of an alkyl group, e.g. methyl, ethyl methylation the addition of a methyl group, usually at lysine or arginine residues. The reverse is called demethylation. isoprenylation or prenylation, the addition of an isoprenoid group (e.g. farnesol and geranylgeraniol) farnesylation geranylgeranylation amidation at C-terminus amino acid addition arginylation, a tRNA-mediation addition polyglutamylation, covalent linkage of glutamic acid residues to tubulin and some other proteins. (See tubulin

polyglutamylase) polyglycation, covalent linkage of one to more than 40 glycine residues to the tubulin C-terminal tail diphthamide formation gamma-carboxylation dependent on Vitamin K glycosylation, the addition of a glycosyl group to either asparagine, hydroxylysine, serine, or threonine, resulting in a glycoprotein. Distinct from glycation, which is regarded as a nonenzymatic attachment of sugars. polysialylation, addition of polysialic acid, PSA, to NCAM glypiation, glycosylphosphatidylinositol (GPI) anchor formation heme moiety may be covalently attached hydroxylation hypusine formation (on conserved lysine of [EIF5A] and aIF5a) iodination (e.g. of thyroid hormones) nucleotides or derivatives thereof may be covalently attached adenylation ADP-ribosylation flavin attachment nitrosylation S-glutathionylation oxidation phosphopantetheinylation, the addition of a 4'-phosphopantetheinyl moiety from coenzyme A, as in fatty acid, polyketide, non-ribosomal peptide and leucine biosynthesis phosphorylation, the addition of a phosphate group, usually to serine, tyrosine, threonine or histidine pyroglutamate formation sulfation, the addition of a sulfate group to a tyrosine. selenylation (co-translational incorporation of selenium in selenoproteins) PTMs involving non-enzymatic additions in vivo glycation, the addition of a sugar molecule to a protein without the controlling action of an enzyme. PTMs involving non-enzymatic additions in vitro biotinylation, acylation of conserved lysine residues with a biotin appendage pegylation PTMs involving addition of other proteins or peptides

ISGylation, the covalent linkage to the ISG15 protein (Interferon-Stimulated Gene 15) SUMOylation, the covalent linkage to the SUMO protein (Small Ubiquitin-related MOdifier) ubiquitination, the covalent linkage to the protein ubiquitin. Neddylation, the covalent linkage to Nedd PTMs involving changing the chemical nature of amino acids

citrullination, or deimination, the conversion of arginine to citrulline deamidation, the conversion of glutamine to glutamic acid or asparagine to aspartic acid eliminylation, the conversion to an

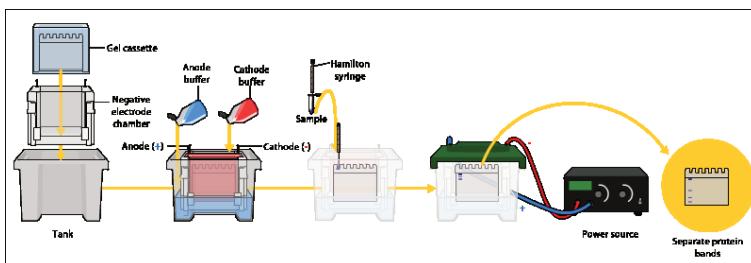
alkene by beta-elimination of phosphothreonine and phosphoserine, or dehydration of threonine and serine, as well as by decarboxylation of cysteine carbamylation, the conversion of lysine to homocitrulline PTMs involving structural changes

disulfide bridges, the covalent linkage of two cysteine amino acids proteolytic cleavage, cleavage of a protein at a peptide bond racemization of proline by prolyl isomerase [18]

Identification of protein in laboratory

The proteins are detected by various method in the laboratory depending upon type of experiment like SDS gel electrophoresis, 2D,Western blotting, massspec etc.

SDS-Gel electrophoresis



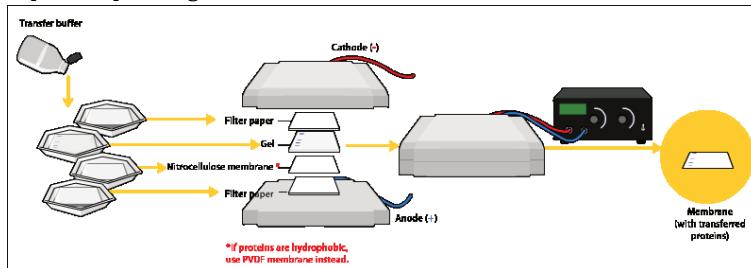
The sample of proteins are separated by using gel electrophoresis. Separation of proteins may be by isoelectric point (pI), molecular weight, electric charge, or a combination of these factors. The nature of the separation depends on the treatment of the sample and the nature of the gel. This is a very useful way to determine a protein. By far the most common type of gel electrophoresis employs polyacrylamide gels and buffers loaded with sodium

dodecyl sulfate (SDS). SDS-PAGE (SDS polyacrylamide gel electrophoresis) maintains polypeptides in a denatured state once they have been treated with strong reducing agents to remove secondary and tertiary structure (e.g. disulfide bonds [S-S] to sulphhydryl groups [SH and SH]) and thus allows separation of proteins by their molecular weight. Sampled proteins become covered in the negatively charged SDS and move to the positively charged electrode through the acrylamide mesh of the gel. Smaller proteins migrate faster through this mesh and the proteins are thus separated according to size (usually measured in kilodaltons, kDa). The concentration of acrylamide determines the resolution of the gel – the greater the acrylamide concentration the better the resolution of lower molecular weight proteins. The lower the acrylamide concentration the better the resolution of higher molecular weight proteins. Proteins travel only in one dimension along the gel for most blots. Samples are loaded into wells in the gel. One lane is usually reserved for a marker or ladder, a commercially available mixture of proteins having defined molecular weights, typically stained so as to form visible, coloured bands. When voltage is applied along the gel, proteins migrate into it at different speeds. These different rates of advancement (different electrophoretic mobilities) separate into bands within each lane.^[19]

Setting of transfer in semi dry assembly

In order to make the proteins accessible to primary and secondary antibody detection, they are moved from within the gel onto a membrane made of nitrocellulose or polyvinylidene difluoride (PVDF). The membrane is placed on top of the gel, and a stack of filter papers placed on top of that. The entire stack is placed in a buffer solution which moves up the paper by capillary action, bringing the proteins with it. Another method for transferring the proteins is called electroblotting and uses an electric current to pull

proteins from the gel into the PVDF or nitrocellulose membrane. The protein move from within the gel onto the membrane while maintaining the organization they had within the gel. As a result of this “blotting” process, the proteins are exposed on a thin surface layer for detection (see below). Both varieties of membrane are chosen for their non-specific protein binding properties (i.e. binds all proteins equally well). Protein binding is based upon hydrophobic interactions, as well as charged interactions between the membrane and protein. Nitrocellulose membranes are cheaper than PVDF, but are far more fragile and do not stand up well to repeated probings.



Some laboratories prefer the wet assembly over the semi dry. The uniformity and overall effectiveness of transfer of protein from the gel to the membrane can be checked by staining the membrane with Coomassie Brilliant Blue or Ponceau S dyes. Ponceau S is the more common of the two, due to Ponceau S's higher sensitivity and its water solubility makes it easier to subsequently destain and probe the membrane.

Facts to be remembered

The central role of proteins as enzymes in living organisms was however not fully appreciated until 1926, when James B. Sumner showed that the enzyme urease was in fact a protein.

The first protein to be sequenced was insulin, by Frederick Sanger, who won the Nobel Prize for this achievement in 1958.

The first protein structures to be solved were hemoglobin and myoglobin, by Max Perutz and Sir John Cowdery Kendrew, respectively, in 1958.

Till 2009, the Protein Data Bank has over 55,000 atomic-resolution structures of proteins.

The first atomic-resolution structures of proteins were solved by X-ray crystallography in the 1960s and by NMR in the 1980s.

Together with Albert Claude and Christian de Duve, George Emil Palade was awarded the Nobel Prize in Physiology or Medicine, in 1974, for the discovery of the ribosomes.

The Nobel Prize in Chemistry 2009 was awarded to Drs Venkatraman Ramakrishnan, Thomas A. Steitz and Ada E. Yonath "for studies of the structure and function of the ribosome."

Question time

1. What do you understand with elongation factors?
2. What are the differences between Eukaryotic and Prokaryotic protein synthesis process?
3. What do you understand with protein synthesis?
4. Define the following terms.
 1. Kilodalton (kDa)
 2. Ribosome
 3. 70S ribosomes vs 80S ribosomes
 4. Kozak consensus
 5. Ribosome Recycling Factor
 6. Peptidyl transfer
5. What is t-RNA and how it is different from m-RNA?
6. Translate the following mRNA codon into one letter and three letter amino acid codes.AUG ACG CGA GCC UGG CCC GGG
GCG CGC AAA ACG GCA GGA ACG ACC AGG UAA

7. Convert the following one letter codes into three letter amino acid codes.A-P-F-G-C-L-K-Q-M-E-H-I-S-V-X

References

1. ↑ Protein
2. ↑ Protein Biosynthesis
3. ↑ Protein
4. ↑ Transfer RNA
5. ↑ Ribosome
6. ↑ http://en.wikipedia.org/wiki/Kozak_consensus_sequence
7. ↑ Malys N, McCarthy JEG (2010). "Translation initiation: variations in the mechanism can be anticipated". *Cellular and Molecular Life Sciences*. doi:10.1007/s00018-010-0588-z.
8. ↑ Structure fo the E. coli protein-conducting channel bound to at translating ribosome, K. Mitra, et al. *Nature* (2005), vol 438, p 318
9. ↑ http://en.wikipedia.org/wiki/Prokaryotic_translation
10. ↑ Hirokawa et al. (2006) "The Ribosome Recycling Step: Consensus or Controversy?". *Trends in Biochemical Sciences* Vol. 31(3), 143-149.
11. ↑ Eukaryotic initiation factor
12. ↑ http://en.wikipedia.org/wiki/Eukaryotic_initiation_factor
13. ↑ http://en.wikipedia.org/wiki/Eukaryotic_initiation_factor
14. ↑ Eukaryotic translation
15. ↑ http://en.wikipedia.org/w/index.php?title=Eukaryotic_translation&oldid=429826292
16. ↑ Protein structure
17. ↑ http://en.wikipedia.org/w/index.php?title=Protein_secondary_structure&oldid=425646887
18. ↑ Posttranslational modification
19. ↑ <http://en.wikipedia.org/w/>

[index.php?title=Western_blot&oldid=429925576](#)

I9.

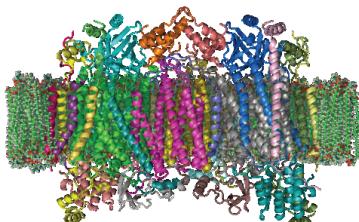
Proteins were first described by the Dutch chemist Gerhardus Johannes Mulder and named by the Swedish chemist Jöns Jakob Berzelius in 1838. Early nutritional scientists such as the German Carl von Voit believed that protein was the most important nutrient for maintaining the structure of the body, because it was generally believed that “flesh makes flesh.”

The amino acids in a polypeptide chain are linked by peptide bonds. Once linked in the protein chain, an individual amino acid is called a residue, and the linked series of carbon, nitrogen, and oxygen atoms are known as the main chain or protein backbone. The peptide bond has two resonance forms that contribute some double-bond character and inhibit rotation around its axis, so that the alpha carbons are roughly coplanar. The other two dihedral angles in the peptide bond determine the local shape assumed by the protein backbone. The end of the protein with a free carboxyl group is known as the C-terminus or carboxy terminus, whereas the end with a free amino group is known as the N-terminus or amino terminus.

The words protein, polypeptide, and peptide are a little ambiguous and can overlap in meaning. Protein is generally used to refer to the complete biological molecule in a stable conformation, whereas peptide is generally reserved for a short amino acid oligomers often lacking a stable three-dimensional structure. However, the boundary between the two is not well defined and usually lies near 20–30 residues. Polypeptide can refer to any single linear chain of amino acids, usually regardless of length, but often implies an absence of a defined Arginineconformation.^[1]

Contents

- 1 Amino acids
 - 1.1 Classification of aminoacids
- 2 Peptide bond
 - 2.1 β -peptides
- 3 Enzymes
 - 3.1 Classification of enzymes
 - 3.2 Oxidoreductase
- 4 Structure of protein
 - 4.1 Primary structure of protein
 - 4.2 Secondary structure of protein
 - 4.3 Tertiary structure of protein
 - 4.4 Quaternary structure of proteins
- 5 Protein structure determination
 - 5.1 X-ray crystallography
 - 5.2 Nuclear magnetic resonance spectroscopy or NMR
- 6 How to sequence a protein?
 - 6.1 Edman degradation
 - 6.2 N-terminal amino acid analysis
 - 6.3 C-terminal amino acid analysis
 - 6.4 Mass spectrometry
- 7 Types of protein
 - 7.1 Conjugated protein
 - 7.1.1 Lipoproteins
 - 7.1.2 Glycoproteins
 - 7.1.3 phosphoproteins
 - 7.1.4 Metalloprotein
 - 7.1.5 hemoproteins
 - 7.1.6 Opsins



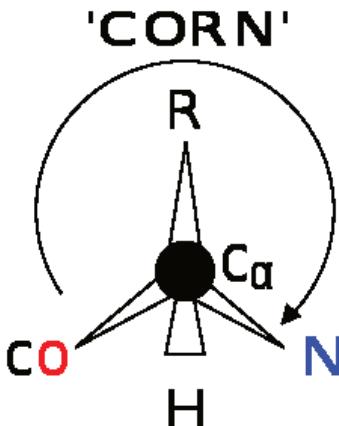
The crystal structure of bovine cytochrome c oxidase in a phospholipid bilayer. The intermembrane space lies to top of the image. PDB 1OCC

- 7.1.7 Flavoproteins
- 7.2 Simple proteins
 - 7.2.1 Albumin
 - 7.2.2 Globulin
 - 7.2.3 Histones
- 7.3 Derived protein
 - 7.3.1 Peptones
 - 7.3.2 Proteases
- 8 Protein data bank or PDB
- 9 Insulin
- 10 References

Amino acids

There are **22 standard** amino acids, but only **21 are found in eukaryotes**. Of the 22, 20 are directly encoded by the universal genetic code. Humans can synthesize 11 of these 20 from each other or from other molecules of intermediary metabolism. The other 9 must be consumed in the diet, and so are called essential amino acids; those are *histidine*, *isoleucine*, *leucine*, *lysine*, *methionine*, *phenylalanine*, *threonine*, *tryptophan*, and *valine*. The remaining two, **selenocysteine** and **pyrrolysine**, are incorporated into proteins by unique synthetic mechanisms.

Each α -amino acid consists of a backbone part that is present in



CO-R-N rule

all the amino acid types, and a side chain that is unique to each type of residue. An exception from this rule is proline, where the hydrogen atom is replaced by a bond to the side chain. Because the carbon atom is bound to four different groups it is chiral, however only one of the isomers occur in biological proteins. Glycine however, is not chiral since its side chain is a hydrogen atom. A simple mnemonic for correct L-form is “CORN”: when the C α atom is viewed with the H in front, the residues read “CO-R-N” in a clockwise direction.^[2]

Isomerism

The standard α -amino acids, all but glycine can exist in either of two optical isomers, called **L or D amino acids**, which are mirror images of each other . While L-amino acids represent all of the amino acids found in proteins during translation in the ribosome, D-amino acids are found in some proteins produced by enzyme posttranslational modifications after translation and translocation to the endoplasmic reticulum, as in exotic sea-dwelling organisms such as cone snails. They are also abundant components of the peptidoglycan cell walls of bacteria, and **D-serine may act as a eurotransmitter in the brain**. The L and D convention for amino acid configuration refers not to the optical activity of the amino acid itself, but rather to the optical activity of the isomer of glyceraldehyde from which that amino acid can theoretically be synthesized (D-glyceraldehyde is dextrorotary; L-glyceraldehyde is levorotary). Alternatively, the (S) and (R) designators are used to indicate the absolute stereochemistry. Almost all of the amino acids in proteins are (S) at the α carbon, with cysteine being (R) and glycine non-chiral.Cysteine is unusual since it has a sulfur atom at the second position in its side-chain, which has a larger atomic mass than the groups attached to the first carbon which is attached to the α -carbon in the other standard amino acids, thus the (R) instead of (S).^[3]

Zwitterions

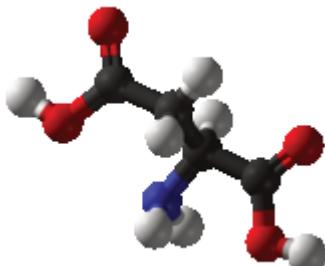
The amine and carboxylic acid functional groups found in amino acids allow it to have amphiprotic properties. At a certain pH, known

as the **isoelectric point**, an amino acid has no overall charge since the number of protonated ammonia groups (positive charges) and deprotonated carboxylate groups (negative charges) are equal. The amino acids all have different **isoelectric points**. The ions produced at the isoelectric point have both positive and negative charges and are known as a zwitterion, which comes from the German word Zwitter meaning “hermaphrodite” or “hybrid”. Amino acids can exist as **zwitterions in solids and in polar solutions such as water**, but **not in the gas** phase. Zwitterions have minimal solubility at their isoelectric point and an amino acid can be isolated by precipitating it from water by adjusting the pH to its particular isoelectric point.^[4]

The 20 naturally occurring amino acids have different physical and chemical properties, including their electrostatic charge, pKa, hydrophobicity, size and specific functional groups. These properties play a major role in molding protein structure. The salient features of amino acids are described below in the table.

Amino Acid	Abbrev.	Remarks
Alanine	A Ala	Very abundant, very versatile. More stiff than glycine, but small enough to pose only small steric limits for the protein conformation. It behaves fairly neutrally, and can be located in both hydrophilic regions on the protein outside and the hydrophobic areas inside.
Asparagine or aspartic acid	B Asx	A placeholder when either amino acid may occupy a position.
Cysteine	C Cys	The sulfur atom bonds readily to heavy metal ions. Under oxidizing conditions, two cysteines can join together in a disulfide bond to form the amino acid cystine. When cystines are part of a protein, insulin for example, the tertiary structure is stabilized, which makes the protein more resistant to denaturation; therefore, disulfide bonds are common in proteins that have to function in harsh environments including digestive enzymes (e.g., pepsin and chymotrypsin) and structural proteins (e.g., keratin). Disulfides are also found in peptides too small to hold a stable shape on their own (eg. insulin).

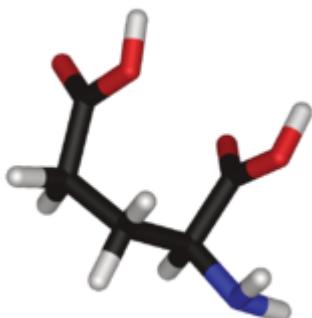
Aspartic acid



D Asp

Behaves similarly to glutamic acid. Carries a hydrophilic acidic group with strong negative charge. Usually is located on the outer surface of the protein, making it water-soluble. Binds to positively-charged molecules and ions, often used in enzymes to fix the metal ion. When located inside of the protein, aspartate and glutamate are usually paired with arginine and lysine.

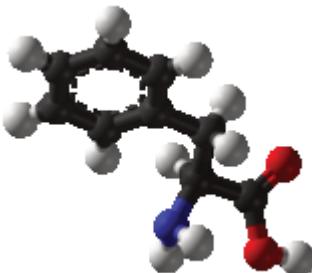
Glutamic acid



E Glu

Behaves similar to aspartic acid. Has longer, slightly more flexible side chain.

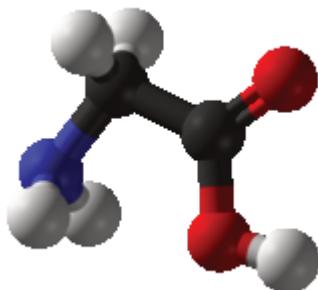
Phenylalanine



F Phe

Essential for humans. Phenylalanine, tyrosine, and tryptophan contain large rigid aromatic group on the side-chain. These are the biggest amino acids. Like isoleucine, leucine and valine, these are hydrophobic and tend to orient towards the interior of the folded protein molecule. Phenylalanine can be converted into Tyrosine.

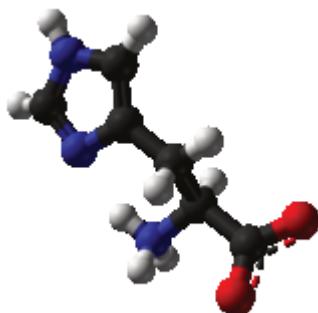
Glycine



G Gly

Because of the two hydrogen atoms at the α carbon, glycine is not optically active. It is the smallest amino acid, rotates easily, adds flexibility to the protein chain. It is able to fit into the tightest spaces, e.g., the triple helix of collagen. As too much flexibility is usually not desired, as a structural component it is less common than alanine.

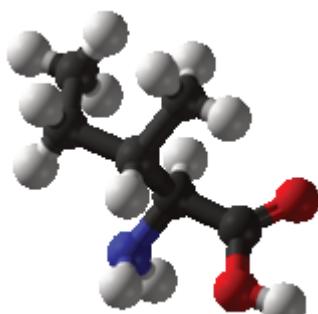
Histidine



H His

In even slightly acidic conditions protonation of the nitrogen occurs, changing the properties of histidine and the polypeptide as a whole. It is used by many proteins as a regulatory mechanism, changing the conformation and behavior of the polypeptide in acidic regions such as the late endosome or lysosome, enforcing conformation change in enzymes. However only a few histidines are needed for this, so it is comparatively scarce.

Isoleucine



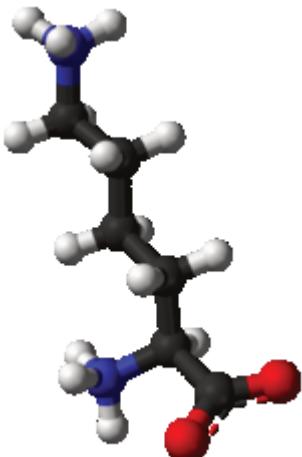
I Ile

Essential for humans. Isoleucine, leucine and valine have large aliphatic hydrophobic side chains. Their molecules are rigid, and their mutual hydrophobic interactions are important for the correct folding of proteins, as these chains tend to be located inside of the protein molecule.

Leucine or isoleucine

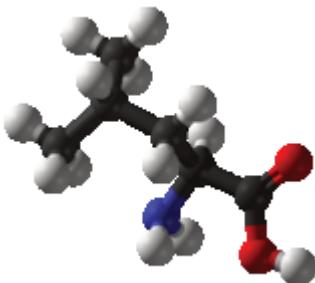
J Xle

A placeholder when either amino acid may occupy a position

Lysine

K Lys

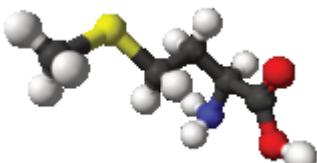
Essential for humans.
Behaves similarly to arginine. Contains a long flexible side-chain with a positively-charged end. The flexibility of the chain makes lysine and arginine suitable for binding to molecules with many negative charges on their surfaces. E.g., DNA-binding proteins have their active regions rich with arginine and lysine. The strong charge makes these two amino acids prone to be located on the outer hydrophilic surfaces of the proteins; when they are found inside, they are usually paired with a corresponding negatively-charged amino acid, e.g., aspartate or glutamate.

Leucine

L Leu

Essential for humans.
Behaves similar to isoleucine and valine. See isoleucine.

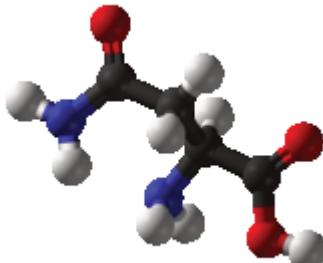
Methionine



M Met

Essential for humans.
Always the first amino acid
to be incorporated into a
protein; sometimes
removed after translation.
Like cysteine, contains
sulfur, but with a methyl
group instead of hydrogen.
This methyl group can be
activated, and is used in
many reactions where a
new carbon atom is being
added to another
molecule.

Asparagine



N Asn

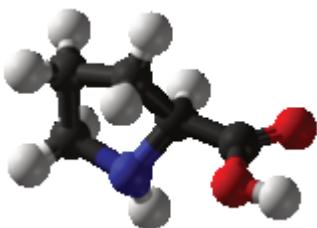
Similar to aspartic acid.
Asn contains an amide
group where Asp has a
carboxyl.

Pyrrolysine

O Pyl

Similar to lysine, with a
pyrrole ring attached.

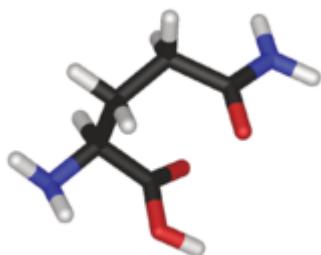
Proline



P Pro

Contains an unusual ring
to the N-end amine group,
which forces the CO-NH
amide sequence into a
fixed conformation. Can
disrupt protein folding
structures like α helix or β
sheet, forcing the desired
kink in the protein chain.
Common in collagen,
where it often undergoes a
posttranslational
modification to
hydroxyproline.

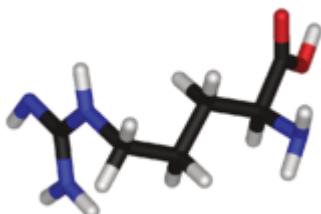
Glutamine



Q Gln

Similar to glutamic acid. Gln contains an amide group where Glu has a carboxyl. Used in proteins and as a storage for ammonia. The most abundant Amino Acid in the body.

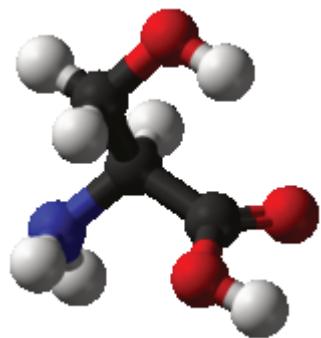
Arginine



R Arg

Functionally similar to lysine.

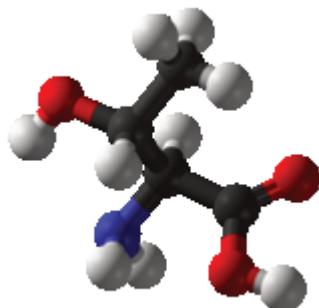
Serine



S Ser

Serine and threonine have a short group ended with a hydroxyl group. Its hydrogen is easy to remove, so serine and threonine often act as hydrogen donors in enzymes. Both are very hydrophilic, therefore the outer regions of soluble proteins tend to be rich with them.

Threonine

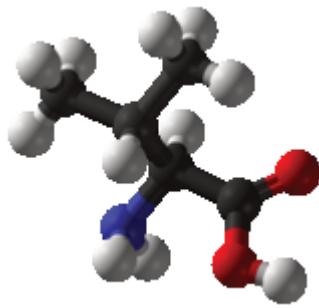


T Thr Essential for humans.
Behaves similarly to serine.

Selenocysteine

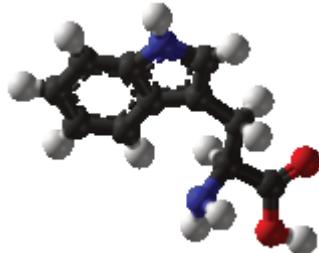
U Sec Selenated form of cysteine, which replaces sulfur.

Valine



V Val Essential for humans.
Behaves similarly to isoleucine and leucine. See isoleucine.

Tryptophan

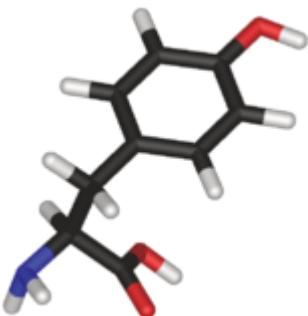


W Trp Essential for humans.
Behaves similarly to phenylalanine and tyrosine (see phenylalanine). Precursor of serotonin. Naturally fluorescent.

Unknown

X Xaa Placeholder when the amino acid is unknown or unimportant.

Tyrosine

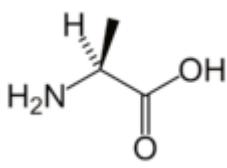


Y Tyr

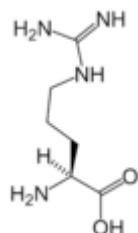
Behaves similarly to phenylalanine (precursor to Tyrosine) and tryptophan (see phenylalanine). Precursor of melanin, epinephrine, and thyroid hormones. Naturally fluorescent, although fluorescence is usually quenched by energy transfer to tryptophans.

Glutamic acid or glutamine

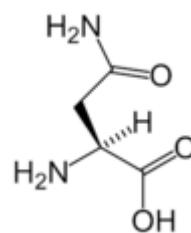
Z Glx A placeholder when either amino acid may occupy a position.



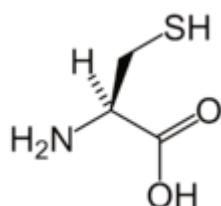
L-Alanine
(Ala / A)



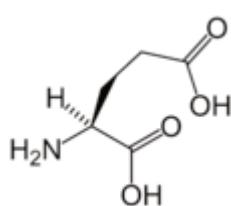
L-Arginine
(Arg / R)



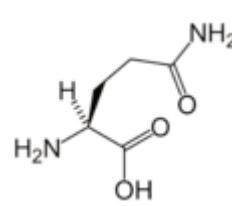
L-Asparagine
(Asn / N)



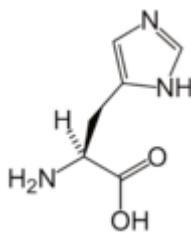
L-Cysteine
(Cys / C)



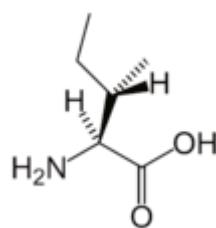
L-Glutamic acid
(Glu / E)



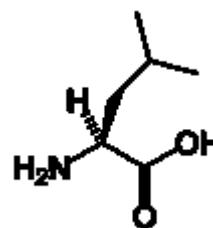
L-Glutamine
(Gln / Q)



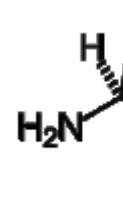
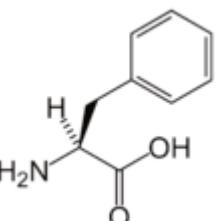
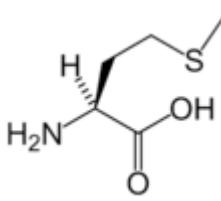
L-Histidine
(His / H)



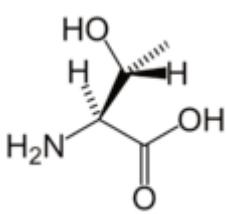
L-Isoleucine
(Ile / I)



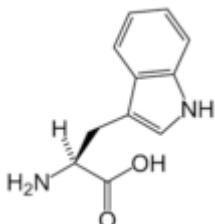
L-Leucine
(Leu / L)



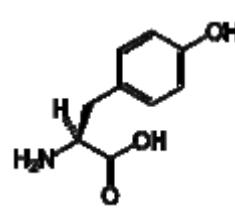
L-Methionine (Met / M)



L-Phenylalanine
(Phe / F)



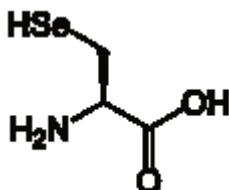
L-Proline (Pro / P)



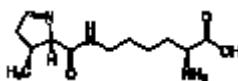
L-Threonine
(Thr / T)



L-Tyrosine (Tyr / Y)



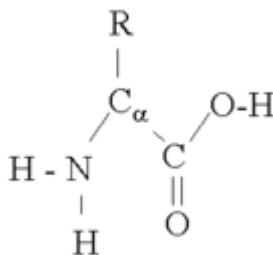
L-Selenocysteine (Sec / U)



L-Pyrrolysine (Pyl / O)

Classification of aminoacids

The 20 amino acids encoded directly by the genetic code can be divided into several groups based on their properties. Important factors are charge, hydrophilicity or hydrophobicity, size and functional groups. Amino acids are usually classified by the properties of their side chain into four groups. The side chain can make an amino acid a weak acid or a weak base, and a hydrophile if the side chain is polar or a hydrophob



An α -amino acid. The $C_\alpha H$ atom is omitted in the diagram.

Protein amino acids are combined into a single polypeptide chain in a condensation reaction. This reaction is catalysed by the ribosome in a process known as translation.

Essential	Nonessential
Isoleucine	Alanine
Leucine	Asparagine
Lysine	Aspartic Acid
Methionine	Cysteine*
Phenylalanine	Glutamic Acid
Threonine	Glutamine*
Tryptophan	Glycine*
Valine	Proline*
	Selenocysteine*
	Serine*
	Tyrosine*
	Arginine*
	Histidine*
	Ornithine*
	Taurine*

Polar and non polar amino acids and their single and three letter code

Amino Acid	Three Letter code	Single Letter code	Side chain polarity	Side chain charge (pH 7.4)	Hydropathy index	Absorbance $\lambda_{\text{max}}(\text{nm})$	$\epsilon_a \lambda_m (\times 10^3 \text{ M}^{-1} \text{ cm}^{-1})$
Alanine	Ala	A	nonpolar	neutral	1.8		
Arginine	Arg	R	polar	positive	-4.5		
Asparagine	Asn	N	polar	neutral	-3.5		
Aspartic acid	Asp	D	polar	negative	-3.5		
Cysteine	Cys	C	nonpolar	neutral	2.5	250	0.3
Glutamic acid	Glu	E	polar	negative	-3.5		
Glutamine	Gln	Q	polar	neutral	-3.5		
Glycine	Gly	G	nonpolar	neutral	-0.4		
Histidine	His	H	polar	positive(10%) neutral(90%)	-3.2	211	5.9
Isoleucine	Ile	I	nonpolar	neutral	4.5		
Leucine	Leu	L	nonpolar	neutral	3.8		
Lysine	Lys	K	polar	positive	-3.9		
Methionine	Met	M	nonpolar	neutral	1.9		
Phenylalanine	Phe	F	nonpolar	neutral	2.8	257, 206, 188	0.2 9.3 60
Proline	Pro	P	nonpolar	neutral	-1.6		
Serine	Ser	S	polar	neutral	-0.8		
Threonine	Thr	T	polar	neutral	-0.7		
Tryptophan	Trp	W	nonpolar	neutral	-0.9	280, 219	5.6 47.
Tyrosine	Tyr	Y	polar	neutral	-1.3	274, 222, 193	1.4 8.0 48.
Valine	Val	V	nonpolar	neutral	4.2		

Additionally, there are two additional amino acids which are incorporated by overriding stop codons:

21st and 22nd amino acids	3-Letter	1-Letter
Selenocysteine	Sec	U
Pyrrolysine	Pyl	O

In addition to the specific amino acid codes, placeholders are used in cases where chemical or crystallographic analysis of a peptide or protein can not conclusively determine the identity of a residue.

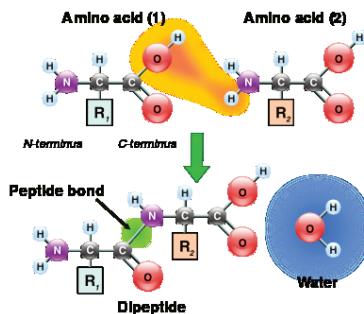
Ambiguous Amino Acids	3-Letter	1-Letter
Asparagine or aspartic acid	Asx	B
Glutamine or glutamic acid	Glx	Z
Leucine or Isoleucine	Xle	J
Unspecified or unknown amino acid	Xaa	X

Unk is sometimes used instead of **Xaa**, but is less standard.

Additionally, many non-standard amino acids have a specific code. For example, several peptide drugs, such as Bortezomib or MG132 are artificially synthesized and retain their protecting groups, which have specific codes. Bortezomib is Pyz-Phe-boroLeu and MG132 is Z-Leu-Leu-Leu-al. Additionally, To aid in the analysis of protein structure, photocrosslinking amino acid analogues are available. These include photoleucine (**pLeu**) and photomethionine (**pMet**).^[5]

Peptide bond

A peptide bond (amide bond) is a covalent chemical bond formed between two molecules when the carboxyl group of one molecule reacts with the amino group of the other molecule, thereby releasing a molecule of water (H_2O). This is a dehydration synthesis reaction (also known as a condensation reaction), and usually occurs between amino acids. The resulting $\text{C}(\text{O})\text{NH}$ bond is called a peptide bond, and the resulting molecule is an amide. The four-atom functional group $-\text{C}(=\text{O})\text{NH}-$ is called a peptide link. Polypeptides and proteins are chains of amino acids held together by peptide bonds, as is the backbone of PNA.



The condensation of two amino acids to form a peptide bond

A peptide bond can be broken by amide hydrolysis (the adding of water). The peptide bonds in proteins are metastable, meaning that in the presence of water they will break spontaneously, releasing 2–4 kcal/mol of free energy, but this process is extremely slow. In living organisms, the process is facilitated by enzymes. Living organisms also employ enzymes to form peptide bonds; this process requires free energy. The wavelength of absorbance for a peptide bond is 190–230 nm.

The peptide bond tends to be planar due to the delocalization of the electrons from the double bond. The rigid peptide dihedral angle, ω (the bond between $\text{C}1$ and N) is always close to 180 degrees. The dihedral angles phi ϕ (the bond between N and $\text{C}\alpha$) and psi ψ (the bond between $\text{C}\alpha$ and $\text{C}1$) can have a certain range of possible values. These angles are the internal degrees of freedom of a protein, they control the protein's conformation. They are restrained by geometry to allowed ranges typical for particular

secondary structure elements, and represented in a Ramachandran plot. A few important bond lengths are given in the table below.^[6]

Peptide bond	Average length	Single bond	Average length	Hydrogen bond	Average (± 30)
Ca – C	153 pm	C – C	154 pm	O-H – O-H	280 pm
C – N	133 pm	C – N	148 pm	N-H – O=C	290 pm
N – Ca	146 pm	C – O	143 pm	O-H – O=C	280 pm

β -peptides

In α amino acids (molecule at left), both the carboxylic acid group (red) and the amino group (blue) are bonded to the same carbon center, termed the α carbon (C^α) because it is one atom away from the carboxylate group. In β amino acids, the amino group is bonded to the β carbon (C^β), which is found in most of the 20 standard amino acids. Only Glycine lacks a β carbon, which means that β -glycine is not possible.

The chemical synthesis of β amino acids can be challenging, especially given the diversity of functional groups bonded to the β carbon and the necessity of maintaining chirality. In the alanine molecule shown, the β carbon is achiral; however, most larger amino acids have a chiral (C^β) atom. A number of synthesis mechanisms have been introduced to efficiently form β amino acids and their derivatives^{[7][8]} notably those based on the Arndt-Eistert synthesis.

Two main types of β -peptides exist: those with the organic residue (R) next to the amine are called β^3 -peptides and those with position next to the carbonyl group are called β^2 -peptides.^[9]

Enzymes

Enzymes are generally globular proteins and range from just 62 amino acid residues in size, for the monomer of 4-oxalocrotonate tautomerase, to over 2,500 residues in the animal fatty acid synthase. A small number of RNA-based biological catalysts exist, with the most common being the ribosome; these are referred to as either RNA-enzymes or ribozymes. The activities of enzymes are determined by their three-dimensional structure. However, although structure does determine function, predicting a novel enzyme's activity just from its structure is a very difficult problem that has not yet been solved.

Most enzymes are much larger than the substrates they act on, and only a small portion of the enzyme (around 3–4 amino acids) is directly involved in catalysis. The region that contains these catalytic residues, binds the substrate, and then carries out the reaction is known as the active site. Enzymes can also contain sites that bind cofactors, which are needed for catalysis. Some enzymes also have binding sites for small molecules, which are often direct or indirect products or substrates of the reaction catalyzed. This binding can serve to increase or decrease the enzyme's activity, providing a means for feedback regulation. Like all proteins, enzymes are long, linear chains of amino acids that fold to produce a three-dimensional product. Each unique amino acid sequence produces a specific structure, which has unique properties. Individual protein chains may sometimes group together to form a protein complex. Most enzymes can be denatured—that is, unfolded and inactivated—by heating or chemical denaturants, which disrupt the three-dimensional structure of the protein. Depending on the enzyme, denaturation may be reversible or irreversible. Structures of enzymes in complex with substrates or substrate analogs during a reaction may be obtained using Time resolved crystallography methods.^[10]

Classification of enzymes

An enzyme's name is often derived from its substrate or the chemical reaction it catalyzes, with the word ending in -ase. Examples are lactase, alcohol dehydrogenase and DNA polymerase. This may result in different enzymes, called isozymes, with the same function having the same basic name. Isoenzymes have a different amino acid sequence and might be distinguished by their optimal pH, kinetic properties or immunologically. Isoenzyme and isozyme are homologous proteins. Furthermore, the normal physiological reaction an enzyme catalyzes may not be the same as under artificial conditions. This can result in the same enzyme being identified with two different names. E.g. Glucose isomerase, used industrially to convert glucose into the sweetener fructose, is a xylose isomerase *in vivo*.

The **International Union of Biochemistry and Molecular Biology** have developed a nomenclature for enzymes, the EC numbers. The **Enzyme Commission number (EC number)** is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze. As a system of enzyme nomenclature, every EC number is associated with a recommended name for the respective enzyme. Each enzyme is described by a sequence of four numbers preceded by "EC". The first number broadly classifies the enzyme based on its mechanism. Strictly speaking, EC numbers do not specify enzymes, but enzyme-catalyzed reactions. If different enzymes (for instance from different organisms) catalyze the same reaction, then they receive the same EC number. By contrast, UniProt identifiers uniquely specify a protein by its amino acid sequence.^[1]

EC 1 Oxidoreductases: catalyze oxidation/reduction reactions

EC 2 Transferases: transfer a functional group (e.g. a methyl or phosphate group)

EC 3 Hydrolases: catalyze the hydrolysis of various bonds

EC 4 Lyases: cleave various bonds by means other than hydrolysis and oxidation

EC 5 Isomerases: catalyze isomerization changes within a single molecule

EC 6 Ligases: join two molecules with covalent bonds.

Top-level EC numbers^[12]

Group	Reaction catalyzed	Typical reaction	Enzyme example(s) with trivial name
EC 1 Oxidoreductases	To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another	AH + B → A + BH (reduced) A + O → AO (oxidized)	Dehydrogenase, oxidase
EC 2 Transferases	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	AB + C → A + BC	Transaminase, kinase
EC 3 Hydrolases	Formation of two products from a substrate by hydrolysis	AB + H ₂ O → AOH + BH	Lipase, amylase, peptidase
EC 4 Lyases	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	RCOCOOH → RCOH + CO ₂ or [x-A-B-Y] → [A=B + X-Y]	Decarboxylase
EC 5 Isomerases	Intramolecule rearrangement, i.e. isomerization changes within a single molecule	AB → BA	Isomerase, mutase
EC 6 Ligases	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP	X + Y + ATP → XY + ADP + Pi	Synthetase

Oxidoreductase

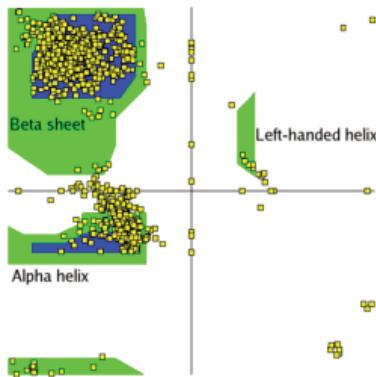
In molecular biology and biochemistry, an oxidoreductase is an enzyme that catalyzes the transfer of electrons from one molecule (the reductant, also called the hydrogen or electron donor) to another (the oxidant, also called the hydrogen or electron acceptor). This group of enzymes usually utilizes NADP or NAD as cofactors. In general, polypeptides are unbranched polymers, so their primary structure can often be specified by the sequence of amino acids along their backbone. However, proteins can become cross-linked, most commonly by disulfide bonds, and the primary structure also requires specifying the cross-linking atoms, e.g., specifying the cysteines involved in the protein's disulfide bonds. Other crosslinks include desmosine... The chiral centers of a polypeptide chain can undergo racemization. In particular, the L-amino acids normally found in proteins can spontaneously isomerize at the C α atom to form D-amino acids, which cannot be cleaved by most proteases.^[13]

Structure of protein

Primary structure of protein

The proposal that proteins were linear chains of α -amino acids was made nearly simultaneously by two scientists at the same conference in 1902, the 74th meeting of the Society of German Scientists and Physicians, held in Karlsbad. **Franz Hofmeister** made the proposal in the morning, based on his observations of the biuret reaction in proteins. Hofmeister was followed a few hours later by **Emil Fischer**, who had amassed a wealth of chemical details supporting the peptide-bond model. For completeness, the proposal that proteins contained amide linkages was made as early as 1882 by the French chemist **E. Grimaux**.^[15]

Despite these data and later evidence that proteolytically digested proteins yielded only oligopeptides, the idea that proteins were linear, unbranched polymers of amino acids was not accepted immediately. Some well-respected scientists such as William Astbury doubted that covalent bonds were strong enough to hold such long molecules together; they feared that thermal agitations would shake such long molecules asunder. Hermann Staudinger faced similar prejudices in the 1920s when he argued that rubber was composed of macromolecules. Thus, several alternative



A Ramachandran plot generated from the protein PCNA, a human DNA clamp protein that is composed of both beta sheets and alpha helices (PDB ID 1AXC). Points that lie on the axes indicate N- and C-terminal residues for each subunit. The green regions show possible angle formations that include glycine, while the blue areas are for formations that don't include glycine.

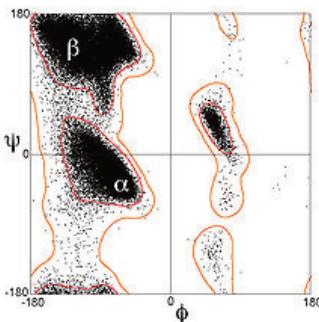
hypotheses arose. The **colloidal protein hypothesis** stated that proteins were colloidal assemblies of smaller molecules. This hypothesis was disproved in the 1920s by ultracentrifugation measurements by Theodor Svedberg that showed that proteins had a well-defined, reproducible molecular weight and by electrophoretic measurements by Arne Tiselius that indicated that proteins were single molecules.

A second hypothesis, the **cyclol hypothesis** advanced by Dorothy Wrinch, proposed that the linear polypeptide underwent a chemical cyclol rearrangement $\text{C=O} + \text{HN C(OH)-N}$ that crosslinked its backbone amide groups, forming a two-dimensional fabric. Other primary structures of proteins were proposed by various researchers, such as the diketopiperazine model of Emil Abderhalden and the pyrrol/piperidine model of Troensegaard in 1942. Although never given much credence, these alternative models were finally disproved when Frederick Sanger successfully sequenced insulin and by the crystallographic determination of myoglobin and hemoglobin by Max Perutz and John Kendrew. The primary structure of peptides and proteins refers to the linear sequence of its amino acid structural units. The term “primary structure” was first coined by Linderstrøm-Lang in 1951. By convention, the primary structure of a protein is reported starting from the amino-terminal (N) end to the carboxyl-terminal (C) end. The post-translational modifications of protein such as disulfide formation, phosphorylations and glycosylations are usually also considered a part of the primary structure, and cannot be read from the gene.

Secondary structure of protein

Secondary structure refers to highly regular local sub-structures. Two main types of secondary structure, the alpha helix and the beta strand, were suggested in 1951 by **Linus Pauling** and coworkers.^[16] These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. They have a regular geometry, being constrained to specific values of the dihedral angles ψ and ϕ on the Ramachandran plot. Both the alpha helix and the beta-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone. Some parts of the protein are ordered but do not form any regular structures. They should not be confused with random coil, an unfolded polypeptide chain lacking any fixed three-dimensional structure. Several sequential secondary structures may form a “supersecondary unit”.^[17]

Amino acids vary in their ability to form the various secondary structure elements. Proline and glycine are sometimes known as “helix breakers” because they disrupt the regularity of the α helical backbone conformation; however, both have unusual conformational abilities and are commonly found in turns. Amino acids that prefer to adopt helical conformations in proteins include methionine, alanine, leucine, glutamate and lysine (“MALEK” in amino-acid 1-letter codes); by contrast, the large aromatic residues (tryptophan, tyrosine and phenylalanine) and C β -branched amino acids (isoleucine, valine, and threonine) prefer to adopt β -strand conformations. However, these preferences are not strong enough



Ramachandran diagram (ϕ,ψ plot), with data points for α -helical residues forming a dense diagonal cluster below and left of center, around the global energy minimum for backbone conformation.^[14]

to produce a reliable method of predicting secondary structure from sequence alone. Secondary structure in proteins consists of local inter-residue interactions mediated by hydrogen bonds, or not. The most common secondary structures are alpha helices and beta sheets. Other helices, such as the 310 helix and π helix, are calculated to have energetically favorable hydrogen-bonding patterns but are rarely if ever observed in natural proteins except at the ends of α helices due to unfavorable backbone packing in the center of the helix.

A Ramachandran plot

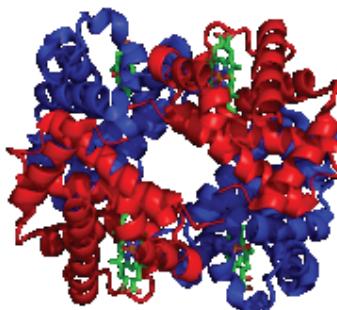
(also known as a Ramachandran map or a Ramachandran diagram or a $[\phi,\psi]$ plot), developed by Gopalasamudram Narayana

Ramachandran and Viswanathan Sasisekharan is a way to visualize dihedral angles ψ against ϕ of amino acid residues in protein structure.^[18] It shows the possible conformations of ψ and ϕ angles for a polypeptide.

Mathematically, the Ramachandran plot is the visualization of a function

$$f: [-\pi, \pi] \times [-\pi, \pi] \rightarrow \mathbb{R}$$

The domain of this function is the torus. Hence, the conventional Ramachandran plot is a projection of the torus on the plane, resulting in a distorted view and the presence of discontinuities. One would expect that larger side chains would result in more restrictions

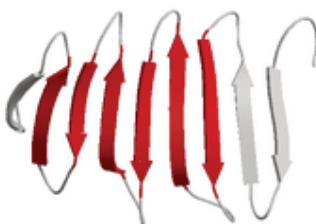


The Hemoglobin molecule has four heme-binding subunits, each largely made of alpha helices.

and consequently a smaller allowable region in the Ramachandran plot. In practice this does not appear to be the case; only the methylene group at the α position has an influence. Glycine has a hydrogen atom, with a smaller van der Waals radius, instead of a methyl group at the α position. Hence it is least restricted and this is apparent in the Ramachandran plot for glycine for which the allowable area is considerably larger. In contrast, the Ramachandran plot for proline shows only a very limited number of possible combinations of ψ and ϕ . The Ramachandran plot was calculated just before the first protein structures at atomic resolution were determined. Forty years later there were tens of thousands of high-resolution protein structures determined by X-ray crystallography and deposited in the Protein Data Bank (PDB). From one thousand different protein chains, Ramachandran plots of over 200 000 amino acids were plotted, showing some significant differences, especially for glycine (Hovmöller et al. 2002). The upper left region was found to be split into two; one to the left containing amino acids in beta sheets and one to the right containing the amino acids in random coil of this conformation. One can also plot the dihedral angles in polysaccharides and other polymers in this fashion. For the first two protein side-chain dihedral angles a similar plot is the Janin Plot.

α helix

The amino acids in an α helix are arranged in a right-handed helical structure where each amino acid residue corresponds to a 100° turn in the helix (i.e., the helix has 3.6 residues per turn), and a translation of 1.5 \AA (0.15 nm) along the helical axis. (Short pieces of left-handed helix sometimes occur with a large content of achiral glycine amino acids, but are unfavorable for the other normal, biological L-amino acids.) The pitch of the alpha-helix (the vertical distance between one consecutive turn of the helix) is 5.4 \AA (0.54 nm) which is the product of 1.5 and 3.6 . What is most important is that the N-H group of an amino acid forms a hydrogen bond with the C=O group of the amino acid four residues earlier; this repeated hydrogen bonding is the most prominent characteristic of an α -helix. Official international nomenclature specifies two ways of defining α -helices, rule 6.2 in terms of repeating ϕ, ψ torsion angles and rule 6.3 in terms of the combined pattern of pitch and hydrogen bonding. Different amino-acid sequences have different propensities for forming α -helical structure. Methionine, alanine, leucine, uncharged glutamate, and lysine ("MALEK" in the amino-acid 1-letter codes) all have especially high helix-forming propensities, whereas proline and glycine have poor helix-forming propensities. Proline either breaks or kinks a helix, both because it cannot donate an amide hydrogen bond (having no amide hydrogen), and also because its sidechain interferes sterically with the backbone of the preceding turn – inside a helix, this forces a bend of about 30° in the helix axis.[9] However, proline is often seen as the first residue of a helix, presumably due to its structural rigidity. At the other extreme, glycine also tends to disrupt helices because its high conformational



Beta-meander motif

Portion of outer surface Protein A of *Borrelia burgdorferi* complexed with a murine monoclonal antibody.

flexibility makes it entropically expensive to adopt the relatively constrained α -helical structure.^[19]

β sheet

The first β sheet structure was proposed by William Astbury in the 1930s. He proposed the idea of hydrogen bonding between the peptide bonds of parallel or antiparallel extended β strands. However, Astbury did not have the necessary data on the bond geometry of the amino acids in order to build accurate models, especially since he did not then know that the peptide bond was planar. A refined version was proposed by Linus Pauling and Robert Corey in 1951.

The β sheet (also β -pleated sheet) is the second form of regular secondary structure in proteins, only somewhat less common than alpha helix. Beta sheets consist of beta strands connected laterally by at least two or three backbone hydrogen bonds, forming a generally twisted, pleated sheet. A beta strand (also β strand) is a stretch of polypeptide chain typically 3 to 10 amino acids long with backbone in an almost fully extended conformation.

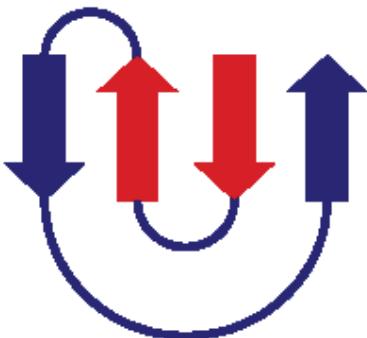
A very simple structural motif involving β sheets is the β hairpin, in which two antiparallel strands are linked by a short loop of two to five residues, of which one is frequently a Glycine or a proline, both of which can assume the unusual dihedral-angle conformations required for a tight turn. However, individual strands can also be linked in more elaborate ways with long loops that may contain alpha helices or even entire protein domains.



Representation of a beta hairpin

Greek key motif

The Greek key motif consists of four adjacent antiparallel strands and their linking loops. It consists of three antiparallel strands connected by hairpins, while the fourth is adjacent to the first and linked to the third by a longer loop. This type of structure forms easily during the protein folding process.^{[20][21]} It was named after a pattern common to Greek ornamental artwork (see meander (art)).



Greek-key motif in protein structure.

The β - α - β motif

Due to the chirality of their component amino acids, all strands exhibit a “right-handed” twist evident in most higher-order β sheet structures. In particular, the linking loop between two parallel strands almost always has a right-handed crossover chirality, which is strongly favored by the inherent twist of the sheet. This linking loop frequently contains a helical region, in which case it is called a β - α - β motif. A closely related motif called a β - α - β - α motif forms the basic component of the most commonly observed protein tertiary structure, the TIM barrel.

β -meander motif

A simple supersecondary protein topology composed of 2 or more consecutive antiparallel β -strands linked together by hairpin loops.^{[22][23]} This motif is common in β -sheets and can be found in several structural architectures including β -barrels and β -propellers.

Psi-loop motif

The psi-loop, Ψ -loop, motif consists of two antiparallel strands with one strand in between that is connected to both by hydrogen bonds.^[24] There are four possible strand topologies for single Ψ -loops as cited by Hutchinson et al. (1990). This motif is rare as the process resulting in its formation seems unlikely to occur during protein folding. The Ψ -loop was first identified in the aspartic protease family.^[25]



Psi-loop motif

Portion of Carboxypeptidase A.

Coiled coils

The possibility of coiled coils for α -keratin was proposed by Francis Crick in 1952 as well as mathematical methods for determining their structure. Remarkably, this was soon after the structure of the alpha helix was suggested in 1951 by Linus Pauling and coworkers.

Coiled coils usually contain a repeated pattern, hxxhcxc, of hydrophobic (h) and charged (c) amino-acid residues, referred to as a heptad repeat. The positions in the heptad repeat are usually labeled abcdefg, where a and d are the hydrophobic positions, often being occupied by isoleucine, leucine or valine. Folding a sequence with this repeating pattern into an alpha-helical secondary structure causes the hydrophobic residues to be presented as a 'stripe' that coils gently around the helix in left-handed fashion, forming an amphipathic structure. The most favorable way for two such helices to arrange themselves in the water-filled environment of the cytoplasm is to wrap the hydrophobic strands against each other sandwiched between the hydrophilic amino acids. It is thus the burial of hydrophobic surfaces, that provides the

thermodynamic driving force for the oligomerization. The packing in a coiled-coil interface is exceptionally tight, with almost complete van der Waals contact between the side chains of the a and d residues. This tight packing was originally predicted by **Francis Crick in 1952** and is referred to as Knobs into holes packing. The α -helices may be parallel or anti-parallel, and usually adopt a left-handed super-coil. Although disfavored, a few right-handed coiled coils have also been observed in nature and in designed proteins.^[26]

Structural features of the three major forms of protein helices^[27]

Geometry attribute	α -helix	β_{10} helix	π -helix
Residues per turn	3.6	3.0	4.4
Translation per residue	1.5 Å	2.0 Å	1.1 Å
Radius of helix	2.3 Å	1.9 Å	2.8 Å
Pitch	5.4 Å	6.0 Å	4.8 Å

Tertiary structure of protein

Tertiary structure is considered to be largely determined by the protein's primary structure – the sequence of amino acids of which it is composed. Efforts to predict tertiary structure from the primary structure are known generally as protein structure prediction. However, the environment in which a protein is synthesized and allowed to fold are significant determinants of its final shape and are usually not directly taken into account by current prediction methods.

In globular proteins, tertiary interactions are frequently stabilized by the sequestration of hydrophobic amino acid residues in the protein core, from which water is excluded, and by the consequent enrichment of charged or hydrophilic residues on the protein's water-exposed surface. In secreted proteins that do not spend time in the cytoplasm, disulfide bonds between cysteine residues help to maintain the protein's tertiary structure. A variety of common and stable tertiary structures appear in a large



The four levels of protein structure, from top to bottom: primary structure, secondary structure (β -sheet left, right α -helix), tertiary and quaternary structure.

number of proteins that are unrelated in both function and evolution – for example, many proteins are shaped like a TIM barrel, named for the enzyme triosephosphateisomerase. Another common structure is a highly stable dimeric coiled coil structure composed of 2-7 alpha helices.

The majority of protein structures known to date have been solved with the experimental technique of X-ray crystallography, which typically provides data of high resolution but provides no time-dependent information on the protein's conformational flexibility. A second common way of solving protein structures uses NMR, which provides somewhat lower-resolution data in general and is limited to relatively small proteins, but can provide time-dependent information about the motion of a protein in solution. Dual polarisation interferometry is a time resolved analytical method for determining the overall conformation and conformational changes in surface captured proteins providing complementary information to these high resolution methods. More is known about the tertiary structural features of soluble globular proteins than about membrane proteins because the latter class is extremely difficult to study using these methods.^[28]

Quaternary structure of proteins

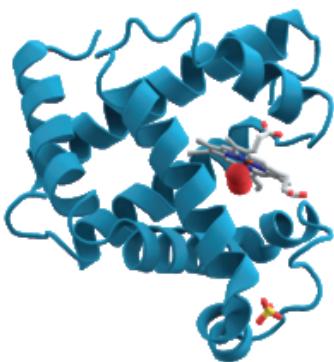
Several proteins are actually assemblies of more than **one polypeptide chain**,

which in the context of the larger assemblage are known as protein subunits. In addition to the tertiary structure of the subunits, multiple-subunit proteins possess a quaternary structure, which is the arrangement into which the subunits assemble. Enzymes composed of subunits with diverse functions are sometimes called holoenzymes, in which some parts may be known as regulatory subunits and the functional core is known as the catalytic subunit. Examples of proteins with quaternary structure include

hemoglobin, DNA polymerase, and ion channels. Other assemblies referred to instead as multiprotein complexes also possess quaternary structure. Examples include **nucleosomes** and **microtubules**.

Changes in quaternary structure can occur through conformational changes within individual subunits or through reorientation of the subunits relative to each other. It is through such changes, which underlie cooperativity and allostery in “multimeric” enzymes, that many proteins undergo regulation and perform their physiological function. The above definition follows a classical approach to biochemistry, established at times when the distinction between a protein and a functional, proteinaceous unit was difficult to elucidate. More recently, people refer to protein-protein interaction when discussing quaternary structure of proteins and consider all assemblies of proteins as protein complexes.^[29]

Protein structure determination



—Ribbon diagram of the structure of myoglobin, showing colored alpha helices. Such proteins are long, linear molecules with thousands of atoms; yet the relative position of each atom has been determined with sub-atomic resolution by X-ray crystallography. Since it is difficult to visualize all the atoms at once, the ribbon shows the rough path of the protein polymer from its N-terminus (blue) to its C-terminus (red).

Around 90% of the protein structures available in the Protein Data Bank have been determined by X-ray crystallography. This method allows one to measure the 3D density distribution of electrons in the protein (in the crystallized state) and thereby infer the 3D coordinates of all the atoms to be determined to a certain resolution. Roughly 9% of the known protein structures have been obtained by Nuclear Magnetic Resonance techniques. The secondary structure composition can be determined via circular dichroism or dual polarisation interferometry.

Cryo-electron microscopy has

recently become a means of determining protein structures to high resolution (less than 5 angstroms or 0.5 nanometer) and is anticipated to increase in power as a tool for high resolution work in the next decade. This technique is still a valuable resource for researchers working with very large protein complexes such as virus coat proteins and amyloid fibers.

X-ray crystallography

X-ray crystallography of biological molecules took off with Dorothy Crowfoot Hodgkin, who solved the structures of cholesterol (1937), vitamin B12 (1945) and penicillin (1954), for which she was awarded the Nobel Prize in Chemistry in 1964. In 1969, she succeeded in solving the structure of insulin, on which she worked for over thirty years.^[30]

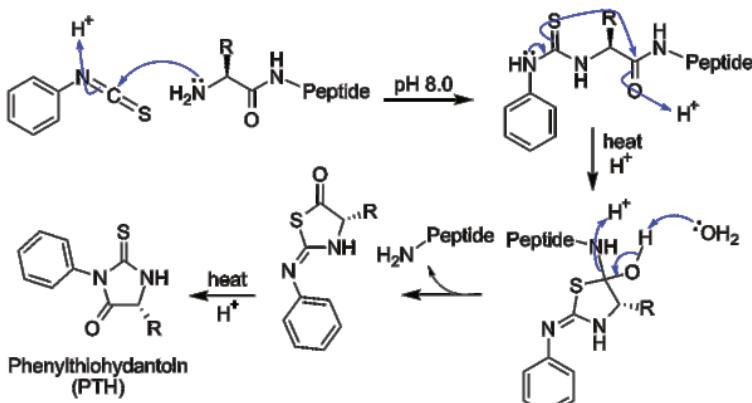
X-ray crystallography is a method of determining the arrangement of atoms within a crystal, in which a beam of X-rays strikes a crystal and diffracts into many specific directions. Crystal structures of proteins (which are irregular and hundreds of times larger than cholesterol) began to be solved in the late 1950s, beginning with the structure of sperm whale myoglobin by Max Perutz and Sir John Cowdery Kendrew, for which they were awarded the Nobel Prize in Chemistry in 1962.^[31] Since that success, over 61840 X-ray crystal structures of proteins, nucleic acids and other biological molecules have been determined.^[32] For comparison, the nearest competing method in terms of structures analyzed is nuclear magnetic resonance (NMR) spectroscopy, which has resolved 8759 chemical structures.^[33] Moreover, crystallography can solve structures of arbitrarily large molecules, whereas solution-state NMR is restricted to relatively small ones (less than 70 kDa). X-ray crystallography is now used routinely by scientists to determine how a pharmaceutical drug interacts with its protein target and what changes might improve it.^[34] However, intrinsic membrane proteins remain challenging to crystallize because they require detergents or other means to solubilize them in isolation, and such detergents often interfere with crystallization. Such membrane proteins are a large component of the genome and include many proteins of great physiological importance, such as ion channels and receptors.^{[35][36]}

Nuclear magnetic resonance spectroscopy or NMR

Protein nuclear magnetic resonance spectroscopy (usually abbreviated protein NMR) is a field of structural biology in which NMR spectroscopy is used to obtain information about the structure and dynamics of proteins. The field was pioneered by Richard R. Ernst and Kurt Wüthrich[1], among others. Protein NMR techniques are continually being used and improved in both academia and the biotech industry. Structure determination by NMR spectroscopy usually consists of several following phases, each using a separate set of highly specialized techniques. The sample is prepared, resonances are assigned, restraints are generated and a structure is calculated and validated

How to sequence a protein?

Protein sequencing is a technique to determine the amino acid sequence of a protein, as well as which conformation the protein adopts and the extent to which it is complexed with any non-peptide molecules. Discovering the structures and functions of proteins in living organisms is an important tool for understanding cellular processes, and allows drugs that target specific metabolic pathways to be invented more easily. The two major direct methods of protein sequencing are **mass spectrometry** and the **Edman degradation reaction**. It is also possible to generate an amino acid sequence from the DNA or mRNA sequence encoding the protein, if this is known. However, there are a number of other reactions which can be used to gain more limited information about protein sequences and can be used as preliminaries to the aforementioned methods of sequencing or to overcome specific inadequacies within them.^[37]



Edman degradation

The Edman degradation is a very important reaction for protein sequencing, because it allows the ordered amino acid composition of a protein to be discovered. Automated Edman sequencers are now in widespread use, and are able to sequence peptides up to approximately 50 amino acids long. A reaction scheme for sequencing a protein by the Edman degradation follows – some of the steps are elaborated on subsequently. Break any disulfide bridges in the protein with an oxidising agent like **performic acid** or reducing agent like **2-mercaptoethanol**. A protecting group such as iodoacetic acid may be necessary to prevent the bonds from reforming. Separate and purify the individual chains of the protein complex, if there are more than one.

Determine the amino acid composition of each chain.

Determine the terminal amino acids of each chain.

Break each chain into fragments under 50 amino acids long.

Separate and purify the fragments.

Determine the sequence of each fragment.

Repeat with a different pattern of cleavage.

Construct the sequence of the overall protein.

Digestion into peptide fragments Peptides longer than about 50-70 amino acids long cannot be sequenced reliably by the Edman degradation. Because of this, long protein chains need to be broken up into small fragments which can then be sequenced individually. Digestion is done either by endopeptidases such as trypsin or pepsin or by chemical reagents such as cyanogen bromide. Different enzymes give different cleavage patterns, and the overlap between fragments can be used to construct an overall sequence.

Phenylisothiocyanate is reacted with an uncharged terminal amino group, under mildly alkaline conditions, to form a cyclical phenylthiocarbamoyl derivative. Then, under acidic conditions, this derivative of the terminal amino acid is cleaved as a thiazolinone derivative. The thiazolinone amino acid is then selectively extracted into an organic solvent and treated with acid to form the more stable phenylthiohydantoin (PTH)- amino acid derivative that can be identified by using chromatography or electrophoresis. This procedure can then be repeated again to identify the next amino acid. A major drawback to this technique is that the peptides being sequenced in this manner cannot have more than 50 to 60 residues (and in practice, under 30). The peptide length is limited due to the cyclical derivitization not always going to completion. The derivitization problem can be resolved by cleaving large peptides into smaller peptides before proceeding with the reaction. It is able to accurately sequence up to 30 amino acids with modern machines capable of over 99% efficiency per amino acid. An advantage of the Edman degradation is that it only uses 10 – 100 picomoles of peptide for the sequencing process. Edman degradation reaction is automated to speed up the process.^{[38][39]}

N-terminal amino acid analysis

Determining which amino acid forms the N-terminus of a peptide chain is useful for two reasons: to aid the ordering of individual

peptide fragments' sequences into a whole chain, and because the first round of Edman degradation is often contaminated by impurities and therefore does not give an accurate determination of the N-terminal amino acid. A generalised method for N-terminal amino acid analysis follows: React the peptide with a reagent which will selectively label the terminal amino acid. Hydrolyse the protein. Determine the amino acid by chromatography and comparison with standards. There are many different reagents which can be used to label terminal amino acids. They all react with amine groups and will therefore also bind to amine groups in the side chains of amino acids such as lysine – for this reason it is necessary to be careful in interpreting chromatograms to ensure that the right spot is chosen. Two of the more common reagents are Sanger's reagent (1-fluoro-2,4-dinitrobenzene) and dansyl derivatives such as dansyl chloride. Phenylisothiocyanate, the reagent for the Edman degradation, can also be used. The same questions apply here as in the determination of amino acid composition, with the exception that no stain is needed, as the reagents produce coloured derivatives and only qualitative analysis is required, so the amino acid does not have to be eluted from the chromatography column, just compared with a standard. Another consideration to take into account is that, since any amine groups will have reacted with the labelling reagent, ion exchange chromatography cannot be used, and thin layer chromatography or high pressure liquid chromatography should be used instead.^[40]

C-terminal amino acid analysis

The number of methods available for C-terminal amino acid analysis is much smaller than the number of available methods of N-terminal analysis. The most common method is to add carboxypeptidases to a solution of the protein, take samples at regular intervals, and

determine the terminal amino acid by analysing a plot of amino acid concentrations against time

Mass spectrometry

Present day researchers are using Mass spectrometry an important tool for the characterization of proteins. **Protein mass spectrometry refers to the application of mass spectrometry to the study of proteins.** The two primary methods for ionization of whole proteins are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). In keeping with the performance and mass range of available mass spectrometers, two approaches are used for characterizing proteins. In the first, intact proteins are ionized by either of the two techniques described above, and then introduced to a mass analyzer. This approach is referred to as “top-down” strategy of protein analysis. In the second, proteins are enzymatically digested into smaller peptides using a protease such as trypsin. Subsequently these peptides are introduced into the mass spectrometer and identified by peptide mass fingerprinting or tandem mass spectrometry. Hence, this latter approach (also called “bottom-up” proteomics) uses identification at the peptide level to infer the existence of proteins.

Whole protein mass analysis is primarily conducted using either time-of-flight (TOF) MS, or **Fourier transform ion cyclotron resonance** (FT-ICR). These two types of instrument are preferable here because of their wide mass range, and in the case of FT-ICR, its high mass accuracy. Mass analysis of proteolytic peptides is a much more popular method of protein characterization, as cheaper instrument designs can be used for characterization. Additionally, sample preparation is easier once whole proteins have been digested into smaller peptide fragments. The most widely used instrument for peptide mass analysis are the MALDI time-of-flight instruments as they permit the acquisition of peptide mass

fingerprints (PMFs) at high pace (1 PMF can be analyzed in approx. 10 sec). Multiple stage quadrupole-time-of-flight and the quadrupole ion trap also find use in this application.

Types of protein

Conjugated protein

A conjugated protein is a protein that functions in interaction with other chemical groups attached by covalent bonds or by weak interactions. Many proteins contain only amino acids and no other chemical groups, and they are called simple proteins. However, other kind of proteins yield, on hydrolysis, some other chemical component in addition to amino acids and they are called conjugated proteins. The nonamino part of a conjugated protein is usually called its prosthetic group. Most prosthetic groups are formed from vitamins. Conjugated proteins are classified on the basis of the chemical nature of their prosthetic groups. Some examples of conjugated proteins are

Lipoproteins

A lipoprotein is a biochemical assembly that contains both **proteins** and **lipids** water-bound to the proteins. Many enzymes, transporters, structural proteins, antigens, adhesins and toxins are lipoproteins. Examples include the high density (HDL) and low density (LDL) lipoproteins which enable fats to be carried in the blood stream, the transmembrane proteins of the mitochondrion and the chloroplast, and bacterial lipoproteins.

Glycoproteins

Glycoproteins are proteins that contain oligosaccharide chains (glycans) covalently attached to polypeptide side-chains. The carbohydrate is attached to the protein in a cotranslational or posttranslational modification. This process is known as glycosylation. In proteins that have segments extending extracellularly, the extracellular segments are often glycosylated. Glycoproteins are often important integral membrane proteins, where they play a role in cell-cell interactions. Glycoproteins also occur in the cytosol, but their functions and the pathways producing these modifications in this compartment are less well-understood. **Glycoproteins are generally the largest and most abundant group of conjugated proteins.** They range from glycoproteins in cell surface membranes that constitute the glycocalyx, to important antibodies produced by leukocytes.

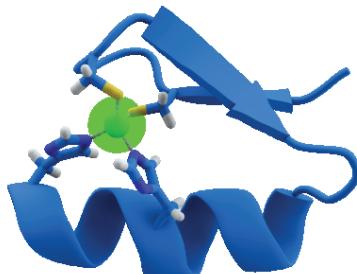
phosphoproteins

Phosphoproteins are proteins which are chemically bonded to a substance containing phosphoric acid (see phosphorylation for more). The category of organic molecules that includes Fc receptors, Ulks, Calcineurins, K chips, and urocortins.

Metalloprotein

A protein that contains a metal ion as a cofactor known as Metalloprotein. Metalloproteins have many different functions in cells, such as enzymes, transport and storage proteins, and signal transduction proteins. Indeed, about one quarter to one third of all proteins require metals to carry out their functions. The metal ion is usually coordinated by nitrogen, oxygen or sulfur atoms belonging

to amino acids in the polypeptide chain and/or a macrocyclic ligand incorporated into the protein. The presence of the metal ion allows metalloenzymes to perform functions such as redox reactions that cannot easily be performed by the limited set of functional groups found in amino acids.



Computer-generated 3-D representation of the zinc finger motif of proteins, consisting of an α helix and an antiparallel β sheet. The zinc ion (green) is coordinated by two histidine residues and two cysteine residues.

Metal Ion	Examples of enzymes containing this ion
Magnesium	Glucose 6-phosphatase Hexokinase DNA polymerase
Vanadium	vanabins
Manganese	Arginase
Iron	Catalase Hydrogenase IRE-BP Aconitase
Nickel ^[41]	Urease Hydrogenase
Copper	Cytochrome oxidase Laccase
Zinc	Alcohol dehydrogenase Carboxypeptidase Aminopeptidase Beta amyloid
Molybdenum	Nitrate reductase
Selenium	Glutathione peroxidase
various	Metallothionein Phosphatase

hemoproteins

A heme protein (or hemoprotein or haemoprotein), or heme protein, is a metalloprotein containing a heme prosthetic group, either covalently or noncovalently bound to the protein itself. The iron in the heme is capable of undergoing oxidation and reduction (usually to +2 and +3, though stabilized Fe+4 and even Fe+5 species are well known in the peroxidases). Hemoproteins probably evolved from a primordial strategy allowing to incorporate the iron (Fe) atom contained within the protoporphyrin IX ring of heme into proteins. This strategy has been maintained throughout evolution as it makes hemoproteins responsive to molecules that can bind divalent iron (Fe). These molecules included, but are probably not restricted to,

gaseous molecules, such as oxygen (O_2) nitric oxide (NO), carbon monoxide (CO) and hydrogen sulfide (H_2S). Once bound to the prosthetic heme groups of hemoproteins these gaseous molecules can modulate the activity/function of those hemoproteins in a way that is said to afford signal transduction. Therefore, when produced in biologic systems (cells), these gaseous molecules are referred to as gasotransmitters. **Haemoglobin contains the prosthetic group containing iron**, which is the haem. It is within the haem group that carries the oxygen molecule through the binding of the oxygen molecule to the iron ion (Fe^{2+}) found in the haem group.^[42]

Hemoglobin

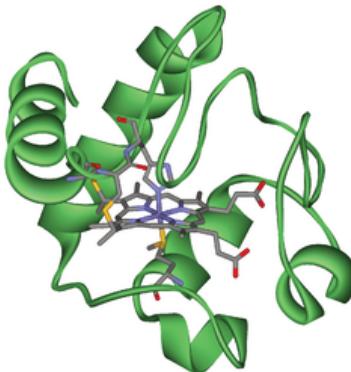
Hemoglobin (also spelled haemoglobin and abbreviated Hb or Hgb) is the iron-containing oxygen-transport metalloprotein in the red blood cells of all vertebrates^[1] (except the fish family Channichthyidae) and the tissues of some invertebrates. Hemoglobin in the blood is what transports oxygen from the lungs or gills to the rest of the body (i.e. the tissues) where it releases the oxygen for cell use, and collects carbon dioxide to bring it back to the lungs. In mammals the protein makes up about 97% of the red blood cells' dry content, and around 35% of the total content (including water)^[citation needed]. Hemoglobin has an oxygen binding capacity of 1.34 ml O_2 per gram of hemoglobin, which increases the total blood oxygen capacity seventyfold. Hemoglobin is involved in the transport of other gases: it carries some of the body's respiratory carbon dioxide (about 10% of the total) as carbaminohemoglobin, in which CO_2 is bound to the globin protein. The molecule also carries the important regulatory molecule nitric oxide bound to a globin protein thiol group, releasing it at the same time as oxygen. Hemoglobin is also found outside red blood cells and their progenitor lines. Other cells that contain hemoglobin include the A9 dopaminergic neurons in the substantia nigra, macrophages, alveolar cells, and mesangial cells in the kidney. In these tissues, hemoglobin has a non-oxygen-carrying function as an antioxidant and a regulator of iron metabolism. Hemoglobin and hemoglobin-like molecules are also found in many invertebrates,

fungi, and plants. In these organisms, hemoglobins may carry oxygen, or they may act to transport and regulate other things such as carbon dioxide, nitric oxide, hydrogen sulfide and sulfide. A variant of the molecule, called leghemoglobin, is used to scavenge oxygen, to keep it from poisoning anaerobic systems, such as nitrogen-fixing nodules of leguminous plants. phytochromes,

Cytochromes

Cytochromes are, in general, membrane-bound

hemoproteins that contain heme groups and carry out electron transport. They are found either as monomeric proteins (e.g., cytochrome c) or as subunits of bigger enzymatic complexes that catalyze redox reactions. They are found in the mitochondrial inner membrane and endoplasmic reticulum of eukaryotes, in the chloroplasts of plants, in photosynthetic microorganisms, and in bacteria.



Cytochrome c with heme c.

Cytochromes Combination

a and a_3	Cytochrome c oxidase ("Complex IV") with electrons delivered to complex by soluble cytochrome c (hence the name)
b and c_1	Coenzyme Q – cytochrome c reductase ("Complex III")
b_6 and f	Plastoquinol–plastocyanin reductase

Type prosthetic group

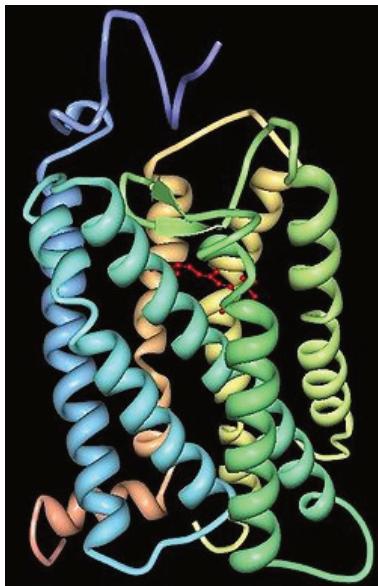
Cytochrome a heme a

Cytochrome b heme b

Cytochrome d tetrapyrrolic chelate of iron

Opsins

Opsins are a group of light-sensitive 35–55 kDa membrane-bound G protein-coupled receptors of the retinylidene protein family found in photoreceptor cells of the retina. Five classical groups of opsins are involved in vision, mediating the conversion of a photon of light into an electrochemical signal, the first step in the visual transduction cascade. Another opsin found in the mammalian retina, melanopsin, is involved in circadian rhythms and pupillary reflex but not in image-forming.



3-dimensional structure of bovine rhodopsin. The seven transmembrane domains are shown in varying colors. The chromophore is shown in red.

Flavoproteins

Flavoproteins are proteins that contain a nucleic acid derivative of riboflavin: the flavin adenine dinucleotide (FAD) or flavin mononucleotide (FMN). Flavoproteins are involved in a wide array of biological processes, including, but by no means limited to, bioluminescence, removal of radicals contributing to oxidative stress, photosynthesis, DNA repair, and apoptosis. The spectroscopic properties of the flavin cofactor make it a natural reporter for changes occurring within the active site; this makes flavoproteins one of the most-studied enzyme families.

Simple proteins

The proteins which upon hydrolysis yield only amino acids are known as simple proteins.

Albumin

Albumin (Latin: albus, white) refers generally to any protein that is water soluble, which is moderately soluble in concentrated salt solutions, and experiences heat denaturation. They are commonly found in blood plasma, and are unique to other plasma proteins in that they are not glycosylated. Substances containing albumin, such as egg white, are called albuminoids.

Globulin

Globulin is one of the three types of serum proteins, the others being albumin and fibrinogen. Some globulins are produced in the liver, while others are made by the immune system. The term globulin encompasses a heterogeneous group of proteins with typical high molecular weight, and both solubility and electrophoretic migration rates lower than for albumin.

Histones

In biology, histones are highly alkaline proteins found in eukaryotic cell nuclei, which package and order the DNA into structural units called nucleosomes. They are the chief protein components of chromatin, acting as spools around which DNA winds, and play a role in gene regulation.

Derived protein

Peptones

Peptones are derived from animal milk or meat digested by proteolytic digestion. In addition to containing small peptides, the resulting spray-dried material includes fats, metals, salts, vitamins and many other biological compounds. Peptone is used in nutrient media for growing bacteria and fungi

Proteases

Proteases occur naturally in all organisms. These enzymes are involved in a multitude of physiological reactions from simple digestion of food proteins to highly-regulated cascades (e.g., the blood-clotting cascade, the complement system, apoptosis pathways, and the invertebrate prophenoloxidase-activating cascade). Proteases can either break specific peptide bonds (limited proteolysis), depending on the amino acid sequence of a protein, or break down a complete peptide to amino acids (unlimited proteolysis). The activity can be a destructive change, abolishing a protein's function or digesting it to its principal components; it can be an activation of a function, or it can be a signal in a signaling pathway.

Protein data bank or PDB

Like fuel and flame, two forces converged to initiate the Protein data bank (PDB): 1) a small but growing data base of sets of protein structures determined by X-ray diffraction and 2) the newly

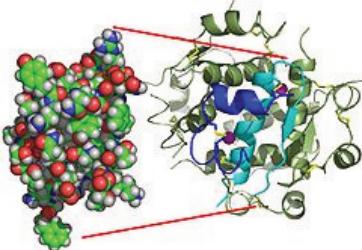
available (1968) molecular graphics display, the BRookhaven Raster Display (BRAD), to inspect these structures in 3-D. In 1969, with the sponsorship of Dr. Walter Hamilton at the Brookhaven National Laboratory, Dr. Edgar Meyer (Texas A&M University) began to write software to store atomic coordinate files in a common format to make them available for geometric and graphical evaluation. By 1971 program SEARCH was executed remotely to extract and examine structural data and thereby was instrumental in initiating networking, thus marking the functional beginning of the PDB.

Upon Hamilton's death in 1973, Dr. Tom Koeztle took over direction of the PDB for the subsequent 20 years. In January 1994, Dr. Joel Sussman of Israel's Weizmann Institute of Science was appointed head of the PDB. In October 1998,^[43] the PDB was transferred to the Research Collaboratory for Structural Bioinformatics (RCSB); the transfer was completed in June 1999. The new director was Dr. Helen M. Berman of Rutgers University (one of the member institutions of the RCSB).^[44] In 2003, with the formation of the wwPDB, the PDB became an international organization. The founding members are PDBe (Europe), RCSB(USA), and PDBj (Japan). The BMRB joined in 2006. Each of the four members of wwPDB can act as deposition, data processing and distribution centers for PDB data. The data processing refers to the fact that wwPDB staff review and annotates each submitted entry. The data are then automatically checked for plausibility (the source code for this validation software has been made available to the public at no charge).

The Protein Data Bank (PDB) is a repository for the 3-D structural data of large biological molecules, such as proteins and nucleic acids. (See also crystallographic database). The data, typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

Insulin

Within vertebrates, the amino acid sequence of insulin is extremely well preserved. Bovine insulin differs from human in only three amino acid residues, and porcine insulin in one. Even insulin from some species of fish is similar enough to human to be clinically effective in humans. Insulin in some invertebrates is quite similar in sequence to human insulin, and has similar physiological effects. The strong homology seen in the insulin sequence of diverse species suggests that it has been conserved across much of animal evolutionary history. The C-peptide of proinsulin, however, differs much more amongst species; it is also a hormone, but a secondary one.



-**The structure of insulin.**

The left side is a space-filling model of the insulin monomer, believed to be biologically active. Carbon is green, hydrogen white, oxygen red, and nitrogen blue. On the right side is a ribbon diagram of the insulin hexamer, believed to be the stored form. A monomer unit is highlighted with the A chain in blue and the B chain in cyan. Yellow denotes disulfide bonds, and magenta spheres are zinc ions.

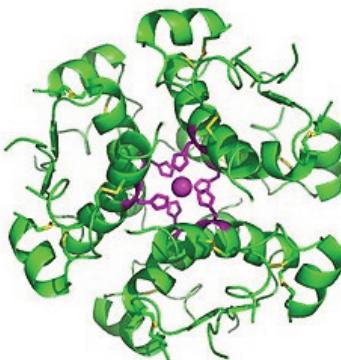
Insulin is produced and stored in the body as a hexamer (a unit of six insulin molecules), while the active form is the monomer. The hexamer is an inactive form with long-term stability, which serves as a way to keep the highly reactive insulin protected, yet readily available. The hexamer-monomer conversion is one of the central aspects of insulin formulations for injection. The hexamer is far more stable than the monomer, which is desirable for practical reasons, however the monomer is a much faster reacting drug because diffusion rate is inversely related to particle size. A fast reacting drug means that insulin injections do not have to precede mealtimes by hours, which in turn gives diabetics more flexibility

in their daily schedule. Insulin can aggregate and form fibrillar interdigitated beta-sheets. This can cause injection amyloidosis, and prevents the storage of insulin for long periods.^[45]

In 1869 Paul Langerhans, a medical student in Berlin, was studying the structure of the pancreas under a microscope when he identified some previously un-noticed tissue clumps scattered throughout the bulk of the pancreas. The function of the “little heaps of cells,” later known as the Islets of Langerhans, was unknown, but Edouard Laguesse later suggested that they might produce secretions that play a regulatory role in digestion. Paul Langerhans’ son, Archibald, also helped to understand this regulatory role. The term insulin originates from insula, the Latin word for islet/island. In 1889, the Polish-German physician Oscar Minkowski in collaboration with Joseph von Mering removed the pancreas from a healthy dog to test its assumed role in digestion. Several days after the dog’s pancreas was removed, Minkowski’s animal keeper noticed a swarm of flies feeding on the dog’s urine. On testing the urine they found that there was sugar in the dog’s urine, establishing for the first time a relationship between the pancreas and diabetes. In 1901, another major step was taken by Eugene Opie, when he clearly established the link between the Islets of Langerhans and diabetes: Diabetes mellitus ... is caused by destruction of the islets of Langerhans and occurs only when these bodies are in part or wholly destroyed. Before his work, the link between the pancreas and diabetes was clear, but not the specific role of the islets.

The Nobel Prize committee in 1923 credited the practical extraction of insulin to a team at the University of Toronto and awarded the Nobel Prize to two men; Fredericus Bantam and J.J.R. Macleod. They were awarded the Nobel Prize in Physiology or Medicine in 1923 for the discovery of insulin. Bantam, insulted that Best was not mentioned, shared his prize with Best, and Macleod immediately shared his with James Collip. The patent for insulin was sold to the University of Toronto for one half-dollar.

The primary structure of insulin was determined by British molecular biologist Frederick Sanger. It was the first protein to have its sequence be determined. He was awarded the 1958 Nobel Prize in Chemistry for this work. In 1969, after decades of work, Dorothy Crowfoot Hodgkin determined the spatial conformation of the molecule, the so-called tertiary structure, by means of X-ray diffraction studies. She had been awarded a Nobel Prize in Chemistry in 1964 for the development of crystallography. Rosalyn Sussman Yalow received the 1977 Nobel Prize in Medicine for the development of the radioimmunoassay for insulin.^[46]



Insulin hexamers highlighting the threefold symmetry, the zinc ions (center) binding with histidine.

References

1. ↑ <http://en.wikipedia.org/w/index.php?title=Protein&oldid=425576197>
2. ↑ http://en.wikipedia.org/w/index.php?title=Proteinogenic_amino_acid&oldid=420804587

3. ↑ http://en.wikipedia.org/w/index.php?title=Amino_acid&oldid=425389108
4. ↑ http://en.wikipedia.org/w/index.php?title=Amino_acid&oldid=425389108
5. ↑ Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. *Nature Methods*:4,261–7,2005
6. ↑ http://en.wikipedia.org/w/index.php?title=Peptide_bond&oldid=417601014
7. ↑ Basler B, Schuster O, Bach T (November 2005). “Conformationally constrained β -amino acid derivatives by intramolecular [2 + 2]-photocycloaddition of a tetrone amide and subsequent lactone ring opening”. *J. Org. Chem.* **70** (24): 9798–808. doi:10.1021/jo0515226. PMID 16292808.
8. ↑ Murray JK, Farooqi B, Sadowsky JD, et al. (September 2005). “Efficient synthesis of a β -peptide combinatorial library with microwave irradiation”. *J. Am. Chem. Soc.* **127** (38): 13271–80. doi:10.1021/ja052733v. PMID 16173757.
9. ↑ Seebach D, Matthews JL (1997). “ β -Peptides: a surprise at every turn”. *Chem. Commun.* (21): 2015–22. doi:10.1039/a704933a.
10. ↑ <http://en.wikipedia.org/w/index.php?title=Enzyme&oldid=424282616>
11. ↑ <http://en.wikipedia.org/w/index.php?title=Enzyme&oldid=424282616>
12. ↑ Moss, G.P.. “Recommendations of the Nomenclature Committee”. International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyse. <http://www.chem.qmul.ac.uk/iubmb/enzyme/>. Retrieved 2006-03-14.
13. ↑ <http://en.wikipedia.org/w/index.php?title=Enzyme&oldid=424282616>
14. ↑ Lovell SC et al. (2003). “Structure validation by C α geometry: φ, ψ and C β deviation”. *Proteins* **50** (3): 437–450. doi:10.1002/

- prot.10286. PMID 12557186.
- 15. ↑ http://en.wikipedia.org/w/index.php?title=Protein_primary_structure&oldid=415921787
 - 16. ↑ Pauling L, Corey RB, Branson HR (1951). "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain". *Proc Natl Acad Sci USA* **37** (4): 205–211. doi:10.1073/pnas.37.4.205. PMID 14816373.
 - 17. ↑ Chiang YS, Gelfand TI, Kister AE, Gelfand IM (2007). "New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage." *Proteins*. **68** (4): 915–921. doi:10.1002/prot.21473. PMID 17557333.
 - 18. ↑ RAMACHANDRAN GN, RAMAKRISHNAN C, SASISEKHARAN V (July 1963). "Stereochemistry of polypeptide chain configurations". *J. Mol. Biol.* **7**: 95–9
 - 19. ↑ http://en.wikipedia.org/w/index.php?title=Alpha_helix&oldid=423162580
 - 20. ↑ Tertiary Protein Structure and Folds: section 4.3.2.1. From Principles of Protein Structure, Comparative Protein Modelling, and Visualisation
 - 21. ↑ Hutchinson EG, Thornton JM (April 1993). "The Greek key motif: extraction, classification and analysis". *Protein Eng.* **6** (3): 233–45. doi:10.1093/protein/6.3.233. PMID 8506258.
 - 22. ↑ SCOP: Fold: WW domain-like
 - 23. ↑ PPS '96 – Super Secondary Structure
 - 24. ↑ Hutchinson, E.; Thornton, J. (1996). "PROMOTIF—A program to identify and analyze structural motifs in proteins". *Protein Science* **5** (2): 212–220. doi:10.1002/pro.5560050204. PMID 8745398.
 - 25. ↑ Hutchinson EG, Thornton JM (1990). "HERA—a program to draw schematic diagrams of protein secondary structures". *Proteins* **8** (3): 203–12. doi:10.1002/prot.340080303. PMID 2281084.
 - 26. ↑ http://en.wikipedia.org/w/index.php?title=Coiled_coil&oldid=427735447

27. ↑ Steven Bottomley (2004). "Interactive Protein Structure Tutorial". <http://www.biomed.curtin.edu.au/biochem/tutorials/prottute/helices.htm>. Retrieved January 9, 2011.
28. ↑ http://en.wikipedia.org/w/index.php?title=Protein_ternary_structure&oldid=422486540
29. ↑ http://en.wikipedia.org/wiki/Protein_quaternary_structure
30. ↑ Crowfoot Hodgkin D (1935). "X-ray Single Crystal Photographs of Insulin". *Nature* **135**: 591. doi:10.1038/135591a0.
31. ↑ Kendrew J. C. et al. (1958-03-08). "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis". *Nature* **181** (4610): 662. doi:10.1038/181662a0. PMID 13517261.
32. ↑ "Table of entries in the PDB, arranged by experimental method". <http://www.rcsb.org/pdb/statistics/holdings.do>.
33. ↑ "PDB Statistics". RCSB Protein Data Bank. http://pdbbeta.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html. Retrieved 2010-02-09.
34. ↑ Scapin G (2006). "Structural biology and drug discovery". *Curr. Pharm. Des.* **12** (17): 2087. doi:10.2174/138161206777585201. PMID 16796557.
35. ↑ Lundstrom K (2006). "Structural genomics for membrane proteins". *Cell. Mol. Life Sci.* **63** (22): 2597. doi:10.1007/s00018-006-6252-y. PMID 17013556.
36. ↑ Lundstrom K (2004). "Structural genomics on membrane proteins: mini review". *Comb. Chem. High Throughput Screen.* **7** (5): 431. PMID 15320710.
37. ↑ http://en.wikipedia.org/w/index.php?title=Protein_sequencing&oldid=413170994
38. ↑ Niall HD (1973). "Automated Edman degradation: the protein sequenator". *Meth. Enzymol.* **27**: 942–1010. doi:10.1016/S0076-6879(73)27039-8. PMID 4773306.
39. ↑ http://en.wikipedia.org/w/index.php?title=Protein_sequencing&oldid=413170994
40. ↑ http://en.wikipedia.org/w/index.php?title=Protein_sequencing&oldid=413170994

41. ↑ Astrid Sigel, Helmut Sigel and Roland K.O. Sigel, ed (2008). *Nickel and Its Surprising Impact in Nature*. Metal Ions in Life Sciences. 2. Wiley. ISBN 978-0-470-01671-8.
42. ↑ <http://en.wikipedia.org/w/index.php?title=Hemeprotein&oldid=410476687>
43. ↑ Berman, H. M.; et al. (January 2000). "The Protein Data Bank". *Nucleic Acids Res.* **28** (1): 235–242. doi:10.1093/nar/28.1.235. PMID 10592235. PMC 102472. <http://nar.oxfordjournals.org/cgi/content/full/28/1/235>.
44. ↑ "RCSB PDB Newsletter Archive". RCSB Protein Data Bank. http://www.rcsb.org/pdb/static.do?p=general_information/news_publications/newsletters/newsletter.html.
45. ↑ <http://en.wikipedia.org/w/index.php?title=Insulin&oldid=425481933>
46. ↑ <http://en.wikipedia.org/w/index.php?title=Insulin&oldid=425481933>

Demo site: https://quizandsurveymaster.com/quiz/sample-quiz/?utm_source=readme&utm_medium=plugin&utm_content=sample-quiz&utm_campaign=qsm_plugin

Plugin site: <https://wordpress.org/plugins/quiz-master-next/>

Quiz No.1[edit]

Point added for a correct answer:

Points for an incorrect answer:

Ignore the questions' coefficients:

1

In DNA Adenine coupled with

T
G
C
W

2

Cytochrome C releases to cytoplasm during

- Apoptosis.
- Mitosis.
- Meiosis.
- Cytosis

3

DNA is a long polymer made from repeating units of

- Nucleotides
- Nucleosides
- Nucleorides
- Nucleophyte

4

How does puromycin inhibit protein synthesis?

- it binds to A site and stop the elongation.
- it affect the peptidyl trasferase activity.
- it degrades the t-rna.
- it is responsible for loading termination codon

5

Micro RNA generally contain

- 22 KB nucleotide
- 22 thousand nucleotide
- 22 nucleotide
- 22 million nucleotide

6

Single letter code for Alanine is

- A
- B
- C
- D

7

The term “Molecular biology” was first used by

- Warren Weaver in 1938
- Warren Buffet
- Watson and Crick
- Gregor Mendel

8

RNAPII can exist in two forms

- Protein and vitamin
- RNAPII0, with a highly phosphorylated CTD, and RNAPIIA, with a nonphosphorylated CTD.
- RNAP 1 and RNAP 2
- Protein and rna

9

DNA replication takes place during

- G1
- G2
- M
- S phase

10

Single letter code for Selenocysteine is

P
B
C
U

11

Single letter code for Phenylalanine is

P
B
C
F

12

What is the main reason for cyanide poisoning?

Inhibition of electron transport chain
Inhibition of Fatty acid transport chain.
Inhibition of Nucleic acid transport chain
Inhibition of Fatty acid metabolism

13

The shape of a DNA molecule is

Double helix.
triple helix.
no helix.
linear

14

Where will you keep collagen triple helix in ramchandran plot

- Top right
- Bottom right
- Bottom left
- Top left

15

Micro RNA was first discovered in

- Yeast
- Bacteria
- C. elegans*
- Mouse

16

Who could not win the Nobel Prize in Physiology or Medicine for discovery of DNA structure

- Franklin
- Watson
- Crick
- Wilkins

17

Single letter code for Tryptophan is

- T
- B
- C
- W

18

Topoisomerases are enzymes with

- both nuclease and ligase activity.
- both polymerase and cutting activity
- both transcription and ligase activity
- both unwinding and polymerase activity

19

t-RNA gene is synthesized by

- RNA polymerase I
- RNA polymerase II
- RNA polymerase III
- DNA polymerase III

20

Who is not associated with discovery of DNA structure

- Linus Pauling
- Watson
- Crick
- Wilkins

Quiz No.2[edit]

Point added for a correct answer:

Points for an incorrect answer:

Ignore the questions' coefficients:

1

Animal cell doesn't contain

- mitochondria
- nucleus
- cell wall
- chromosome

2

Ocher codon is

- UAG
- UGA
- UAA
- UUU

3

The term CDKs stands for

- cyclin direct kinases
- cyclin depen kinases
- cyclin-dependent kinases
- cyclin-dull kinases

4

In cell cycle term “cdc” stands for

- cyclin direct cycle
- cell dependent cycle
- cell division cycle
- cell-dull cycle

5

DNA is a double helix (if true then explain if false then explain)

- TRUE.
- FALSE.

6

m-RNA is a single stranded structure (if true then explain if false then explain)

- TRUE.
- FALSE.

7

Translation of protein generally occurs in the nucleus, (if true then why and if false then why)

- TRUE.
- FALSE.

8

tRNA appears like cloverleaf structure (if true then why and if false then why)

- TRUE.
- FALSE.

9

The first topoisomerase was discovered in

Yeast
Human
E. coli
Mouse

10

The first topoisomerase was discovered by

James Watson
James Xang
James C. Wang
Crick

11

Dideoxynucleotide Chain-termination methods was discovered by

James Watson
James Xang
Frederick Sanger
Crick

12

PCR methods typically amplify DNA fragments of up to

10 bp
100 bp
10 kb
10 mb

Note[edit]

Dear Readers

If you have problem regarding any question in quiz please let me know.

Thanks

with regards

kaushlendra tripathi

Retrieved from "https://en.wikibooks.org/w/index.php?title=An_Introduction_to_Molecular_Biology/Quiz_time&oldid=3251502"

Category:

- Book:An Introduction to Molecular Biology

Alanine

Alanine (abbreviated as Ala or A) is an α -amino acid with the chemical formula $\text{CH}_3\text{CH}(\text{NH}_2)\text{COOH}$. The L-isomer is one of the 22 proteinogenic amino acids.

Alpha helix

A common motif in the secondary structure of proteins, the alpha helix (α -helix) is a right-handed coiled or spiral conformation, in which every backbone N-H group donates a hydrogen bond to the backbone C=O group of the amino acid four residues earlier.

Aminoacids

Amino acids are molecules containing an amine group, a carboxylic acid group and a side chain that varies between different amino acids. The key elements of an amino acid are carbon, hydrogen, oxygen, and nitrogen.

Aliphatic compounds

Aliphatic compounds are acyclic or cyclic, non-aromatic carbon compounds. Thus, aliphatic compounds are opposite to aromatic compounds.

Asparagine

Asparagine (abbreviated as Asn or N) is one of the 20 most common natural amino acids on Earth. It has carboxamide as the side chain's functional group. It is not an essential amino acid.