Central dogma of molecular biology

The term "central dogma of molecular biology" is patterned after religious terminology. However, it refers to a process that is subject to the changes in understanding that are associated with any scientific research. The most simplified form of the central dogma is that the flow of information is from DNA □ RNA □ Protein. This concept has been subject to alterations as our understanding of the processes involved has changed.

The processes involved are transcription and translation. It was understood at the time that some modifications to this pathway were necessary. The most obvious is that DNA is used as the template for DNA replication.

More recently, RNA viruses, in which DNA is never involved in the life cycle, have been discovered. Some of these are retroviruses, in which RNA is used as a template for DNA synthesis in a process called "reverse transcription".

Other modifications to the simple scheme of information flow are proteins act as gene transcriptional regulators, the discovery that some information is stored in methylation patterns of the DNA, and the discovery of prions (proteins that can transmit information to other proteins).

In addition, the existence of introns and exons means that the information stored in the DNA is not always reflected in the mRNA and protein products. A **gene** is stretch of DNA containing both a template for RNA synthesis and sequences that allow the control of RNA production from the template region. However, in many cases, more than one protein can be produced from a DNA sequence, and the coding sequence is not necessarily linearly contiguous within the DNA.

More unusual exceptions to the central dogma include the process for expressing the ApoB gene in humans. The human liver expresses the full length ApoB protein; however, the human intestines mutate a C to U in the ApoB *mRNA* to create a stop codon, and therefore synthesize a shorter protein product although the DNA is not affected. Trypanosomes (the parasitic organism responsible for sleeping sickness) insert additional U nucleotides into some of their mRNA to produce proteins that are not directly coded by the DNA.

How big is DNA?

The amount of DNA required to provide the genetic information for an organism varies fairly dramatically depending on the organism. DNA molecules are usually described in terms of the number of paired monomer units in the double stranded helix, with these units called base-pairs (abbreviated bp). Because DNA molecules tend to be large, larger units, such as kb, for kilobase pairs, or Mb for megabase pairs are also frequently used. More recently, the term Gb, for gigabase pair has been used to refer to the cellular DNA content of higher eukaryotes.

Genome sizes vary over a wide range. Viruses tend to have small genomes, from 5 to 200 kb, but viruses are not free living organisms. Other genomes are listed below. Organisms with gene counts listed have had their genomes sequenced, although some of the genome projects are incomplete. The number of genes and the genome sizes for these organisms are subject to revision as more information becomes available.

Note that genome size is not necessarily related to complexity of the organism; the organism with the largest genome in this table is a single-celled organism. However, prokaryotes always have smaller genomes than eukaryotes (the reasons for this are discussed below in the section on replication).

Species	Type of organism	Genome size (bp)	Genes
Haemophilus influenzae	pathogenic bacterium	1,830,138	1740
Escherichia coli	enteric bacterium	4,639,221	4,377
Escherichia coli O157:H7	pathogenic variant	5,440,000	5,416
Saccharomyces cerevisiae	baker's yeast	12,067,280	6034
Arabidopsis thaliana	smallest plant genome (the plant is a weed)	117,000,000	25,498
Drosophila melanogaster	fruit fly	180,000,000	13,061
Caenorhabditis elegans	nematode worm	100,000,000	19,820
Gallus gallus	chicken	1,200,000,000	?
Mus musculus	mouse	3,454,200,000	?
Pan troglodytes	chimpanzee	3,600,000,000	?
Homo sapiens	human	3,400,000,000	30,000 to 45,000
Pinus resinosa	pine tree	68,000,000,000	?
Amoeba dubia	amoeba	670,000,000,000	?

Note: for most of the eukaryotic organisms, the genome size list corresponds to the haploid genome; most eukaryotic cells are diploid and have twice this amount of nuclear DNA. The amoeba may have a polyploid genome and probably has a smaller amount of unique DNA sequence.

Humans have 46 chromosomes, and therefore the size of the average human chromosome is ~145 million base pairs. Clearly these molecules are much larger than the chromosomes from the bacterial organisms.

DNA molecules are the largest biological molecules. A very large protein has a molecular weight of $\sim 10^6$. By comparison the *E. coli* chromosome, a moderately small DNA molecule, has a molecular weight of ~3 x 10^9.

The process of DNA synthesis

Proper replication requires a large number of different enzymes. One obvious enzyme is the DNA polymerase that actually incorporates the new DNA molecules. Most organisms have a number of DNA polymerases. The majority of these enzymes are used to proofread the newly synthesized DNA, and to repair mistakes incorporated during synthesis or as a result of damage to the DNA. One of the DNA polymerase types, called DNA polymerase III in *E. coli*, is the specialized replication polymerase. The replication polymerase is highly **processive**: it is capable of synthesizing >500,000 bases without dissociating from the template.

In order for a DNA polymerase to synthesize a new DNA strand, the two strands of the double helix must be separated. This process requires energy – an enzyme, *DNA helicase*, unwinds the strands, expending 2 ATP per base pair. The DNA polymerase then adds the next dNTP (using the template strand to ensure specificity of dNTP selection), and continues on.

The polymerization process seems simple. In practice, however, a significant problem exists: the DNA has two strands that run in opposite directions, while all DNA polymerases synthesize DNA in the 5´ to 3´ direction. For one strand, this is readily accomplished; for the other, the synthesis must occur in steps, because the same replication DNA polymerase complex synthesizes both strands.

In order to replicate the opposite strand (usually termed the lagging strand), the polymerase complex must loop the DNA. Its ability to do so is limited by the fact that the DNA must be unwound and by other considerations. In addition, DNA polymerases require a starting place. A short RNA primer supplies this necessary starting place.

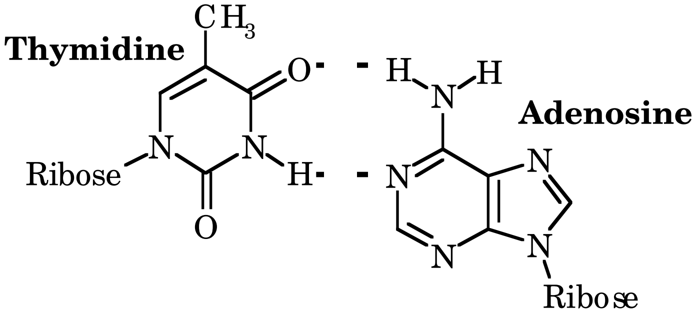
General procedure of polymerization

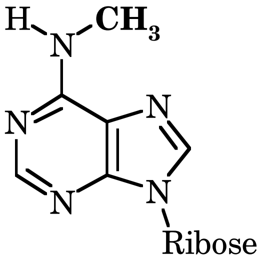
Leading strand synthesis is continuous process.
Lagging strand synthesis has several steps: the synthesis results in formation of many short stretches of new DNA that must be later connected together. The short stretches of DNA are called **Okazaki fragments**.
1) Enzymes induce a local strand separation.
2) A specific RNA polymerase (sometimes called a primase) synthesizes an RNA primer.
3) The DNA polymerase starts at the RNA primer and synthesizes DNA until it reaches the end of the previous fragment.
4) DNA ligase forms a covalent bond between the last base of the new fragment and the first base of the previous fragment.
5) RNase H degrades the RNA primer in preparation for synthesis of the next fragment.

Mistakes

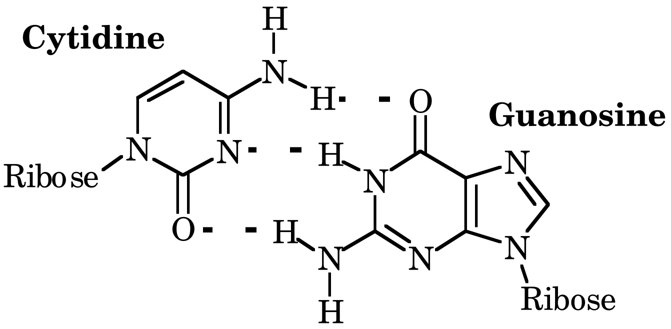
The replication DNA polymerase is highly specific, but occasionally it adds the incorrect base (this is more common if the cell contains excessive amounts of one dNTP during DNA synthesis). The incorporation error rate of most replication polymerases is about 1 in 10^4 to 10^5 bases added. Polymerases can also “stutter” by putting in additional bases that do not base pair to the template, or by leaving out one or more bases (these artifacts are more common in regions where the sequence has strings of one base).

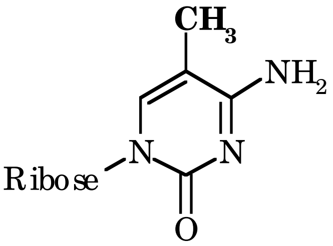
Mistakes can be corrected in several ways. The first method is based on an activity inherent in the polymerase complex. The polymerase complex has proofreading functions that can remove the last base added, and replace it with the correct one. This immediate 3´-5´ exonuclease function reduces the error rate by a factor of 10^2 to 10^3. In addition, proofreading enzymes patrol the DNA looking for mistakes in the newly synthesized strand. Note that the two strands of the double helix are **not** identical. This is because the old strand has methyl groups added to some of the bases, predominantly the N^6 of adenines and the 5-carbon of cytidines. The methylation allows the DNA synthesis machinery to differentiate the old strand and new strands. The methylation occurs at positions in these bases that does not interfere with proper base pairing.





N6-methyladenosine





5-methylcytidine

Topoisomerases

DNA coiling must be controlled. Relaxed DNA forms the standard double helix. Separating the strands tightens the coiling of the DNA that has not yet been replicated. In addition, the cell often needs to change the coiling of the DNA. Replication of circular DNA results in interlocking rings, which must be separated in order to allow the daughter cells to each have a copy. Finally, DNA can “tangle” – anyone ever working with thread or string knows about tangles, and DNA molecules are, in effect, very, very long strings.

These problems are solved by *topoisomerases*, enzyme that can cleave DNA and alter the tightness of the winding, and can pass one strand through another.

Eukaryotic DNA replication

Eukaryotic cells are much more complex than prokaryotic cells, and contain much more DNA. Human cells contain more than 1000 times more DNA than *E. coli*. The polymerase responsible for human DNA replication is about 10-fold slower than the *E. coli* enzyme.

Replication must copy the ***entire*** *genome* ***once*** *and only once.* Prokaryotic organisms manage this by having a single replication origin on each DNA molecule. However, to allow replication to occur in a reasonable amount of time (*i.e.* the 4-5 hours of S phase), eukaryotes must simultaneously replicate DNA at many locations in their genome.

Cell cycle control proteins coordinate the initiation of replication; this still leaves the potential problem of missing some regions. One mechanism for checking is DNA methylation, but the precise mechanism that the cell uses to copy every base exactly one time for each cell division is incompletely understood.

The actual replication process (*i.e.* DNA strand separation and polymerization, with a leading and lagging strand) is similar in most respects to the process used in prokaryotic cells. The enzymes are somewhat different, but eukaryotes use the same general procedure.

Humans do not have exact analogs of the *E. coli* DNA polymerase III; the mammalian replication polymerase is a large complex with at least three separate polymerases: pol □, pol □, and pol □. Pol □ is the major highly processive polymerase, but all three proteins are involved in mammalian DNA replication.

Structural analysis of the DNA replication polymerases indicates that the polymerases form a circular structure that completely surrounds the DNA strand. This structure prevents the polymerase from dissociating from the DNA unless the DNA strand is broken or the polymerase complex is disrupted.

Telomeres

A major difference between prokaryotic and eukaryotic replication occurs at the ends of the chromosomes. In contrast to circular prokaryotic genomes, eukaryotic chromosomes are linear molecules. This means that it is difficult to synthesize the lagging strand at the end of the chromosome, because there is no place to put the primer required to initiate Okazaki fragment synthesis. This means that each eukaryotic cell division results in chromosomes that are slightly (50-100 bp) shorter than those in the parent cell).

The ends of the chromosomes are called **telomeres**; telomeres are repeats of 6 bp sequences of DNA (TTAGGG). At least in part as a result of the shortening of the telomeres, most eukaryotic cells are only capable of dividing a limited number of times. Eventually, the telomeres become too short, and the cell is no longer capable of cell division. (This is not, however, the only method for controlling cessation of cell division.)

Some cells must continue dividing indefinitely. These cells avoid problems with telomere shortening by expressing an enzyme called telomerase. **Telomerase** is a reverse transcriptase-like enzyme. It contains both RNA and protein subunits; the RNA acts as a template for the synthesis of the telomere 6 bp repeat, while the protein contains the catalytic activity. Telomerase is capable of lengthening the telomeres, and therefore preventing damage to the ends of the chromosomes.

Telomerase is required for stem cells, and for germline cells. In addition most (although not all) cells express telomerase as part of their journey towards becoming tumors.

Note, however, that most cells have telomeres longer than are actually necessary:

Number of cell divisions	Approximate mass of progeny of single cell (kilograms)
40	1
50	1000
60	1,000,000

In another illustration of this point, mice with telomerase deleted were able to reproduce for 6 generations (the final generation was infertile). Clearly, most organisms have telomeres longer than is necessary for normal cell division.

DNA damage

DNA is special. Unlike other biological molecules, which are degraded if they become non-functional due to damage or to mistakes during synthesis, DNA must be repaired. DNA repair is a constant process; each human cell suffers DNA damage about 10,000 times each day.

Damage can occur as the result of chemical reactions involving the bases, as a result of light (usually ultraviolet light) induced chemical changes in the bases, as a result of ionizing radiation, or as a result of radioactive decay of atoms in the DNA molecule itself. (With ~100 billion total atoms in human chromosomal DNA, it is likely that some nucleotides will contain radioactive isotopes of the standard chemical elements).

Damage repair differs from mistake correction: mistakes are the result of incorrect base addition during synthesis, and are corrected based on the older of the two strands (usually determined by the presence of methylation on the older strand). In contrast, damage repair must use the other strand, because no other source of the "correct" sequence exists.

Most damage repair involves the removal of a small segment of one strand, followed by the action of a DNA polymerase, using the other strand as a template. Repair fidelity is generally very high (eukaryotes incorporate only one mistake in $\sim 10^{10}$ bases repaired). However, this means that mutations are quite common in organisms with large genomes.

Why don't all humans die of the consequences of mutations? The answer is complicated (and not fully understood).

Part of the reason is that the vast majority of mammalian DNA (more than ~95%) does not code for proteins and is not involved in major regulatory processes. This non-coding DNA has been termed "junk DNA" (with "junk" being something useless that is kept, in contrast to "garbage", which is discarded). This junk DNA includes DNA that comprises the introns and DNA that resides in between genes. In addition, in multicellular organisms, most cells express only a small subset of genes; damage to non-expressed genes will usually not affect these cells.

Another part of the reason is that mutations that occur within important regions may kill one cell, but that cell can (at least in most tissues) be readily replaced. In addition, many cells have redundant pathways; if one gene is inactivated, others may be able to substitute for it.