Biological Background

Amar Mukherjee
School of Computer Science
University of Central Florida, Orlando
Email: amar@cs.ucf.edu

Genetics and Molecular Biology

Genes are discrete physical entities present in all living organisms that control hereditary characteristics passed from parents to offspring of organisms.

Study of heredity is called **Genetics**. Gregor Mendel an unknown monk in Brno (now in Czechoslovakia, it was in Austria) published a paper in 1886 that pioneered the experimental study of genetics and his famous laws are called **Mendel's laws**. Classical genetics assumed genes are abstract attributes occurring in variant forms (called *alleles*). Each individual inherits two genes, one from each of its parents.

In the 1930, it was recognized that like all particles in human body, genes must be composed of molecules and the field devoted to understanding the chemical nature of genes was termed **molecular biology**.

Biological Information

Soon biologists realized that genes are not merely units of inheritance but they are actually units of **biological information** that control all aspects of life –birth, growth, functioning as a living organism and death.

Bioinformatics and Computational Molecular Biology are concerned with the use of computing and mathematical sciences as tools to advance traditional laboratory-based biology.

The need to process an exponentially growing amount of biological information for further scientific advances and to understand its role in heredity, chemical processes within the cell, drug discovery, evolutionary studies etc. have created new problems that are of interdisciplinary nature.

Protein or DNA

- By 1920s, it was established that:
  - Genes reside on chromosomes
  - Chromosomes are made of protein and DNA

  Chromosomes were discovered in 19th century as threadlike structures in the nucleus of a eukaryotic cell that could be observed under microscope as the cells begin to divide.

  Biochemical analysis concluded that chromosomes contain both DNA and protein.
- The question was: What is the genetic material?
  - Protein?
  - DNA?
  - Both?

  The chemical structures of both DNA and protein were still unknown mysteries.

Properties of Genetic Material

- Genetic Material:
  - Must be able to exist in almost infinite variety of forms
  - Protein was believed to be able to form long chains – **macro molecules**. Many proteins were known.
  - DNA was believed to be a small, invariant molecule.

From the point of view of variability in species, proteins as carriers of genetic information seemed to make more sense. But this hypothesis was proved to be false by a famous experiment (by Griffith in 1942 and interpreted by Avery) that discovered a fundamental principle of Biology, called the **Transforming Principle.**

The Transforming Principle

- *Diplococcus pneumoniae* exists in two forms:
  - Both the forms have a coating that surrounds the cell
  - The coating is made of a *polysaccharide* secreted by the bacterium
  - Each form secretes a different polysaccharide, hence a different coating & appearance.
  - The *smooth* (or *S*) form is virulent, whereas the *rough* (or *R*) form is avirulent.

Experiment (Griffith,Avery 1928)

- Mouse injected with S form -> infected with pneumonia
- Mouse injected with R form -> healthy
- Mouse injected with heat-killed S form -> healthy
- Mouse injected with heat-killed S bacteria + live R bacteria -> infected with pneumonia
- Conclusion – a component (the genetic material) of heat-killed S-bacteria was able to enter R cell and transform them into smooth form.

Eukaryote and Prokaryote

Prokaryotes are the so-called ‘lower’ organisms which lack extensive cellular structures viz. membranes, organelles and the genetic material is not enclosed in a ‘nucleous’.

Eukaryotes are organisms composed of one or many cells, each cell having a **nucleous** and a **cytoplasm**. This covers all living organisms except viruses and prokaryotes (bacteria and archea).

Eukaryotes and prokaryotes diverged at the early stages of cell evolution.

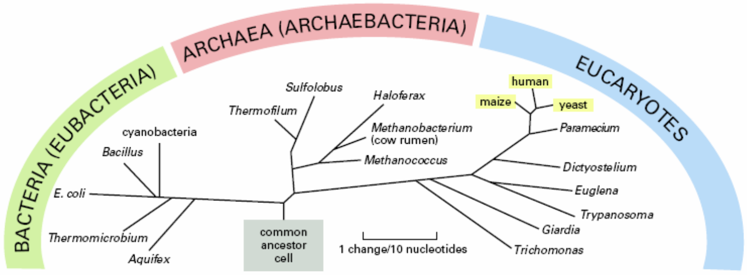
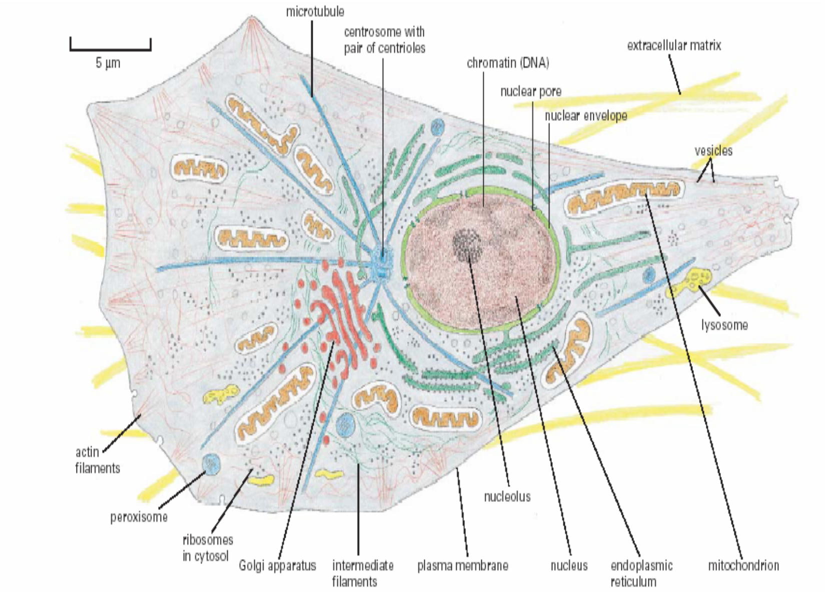


Figure 1–21 The three major divisions (domains) of the living world. Note that traditionally the word *bacteria* has been used to refer to procaryotes in general, but more recently has been redefined to refer to eubacteria specifically. Where there might be ambiguity, we use the term *eubacteria* when the narrow meaning is intended. The tree is based on comparisons of the nucleotide sequence of a ribosomal RNA subunit in the different species. The lengths of the lines represent the numbers of evolutionary changes that have occurred in this molecule in each lineage (see Figure 1–22).

```
GTTCCGGGGGGAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCC  human
|   |  |||||||| ||| ||||  ||||||||||||||||||| |||  | |||| || | | |||||||||||||| |||||| || ||||| | ||  | || ||| 
GCCGCCTGGGGAGTACGGTCGCAAGACTGAAACTTAAAGGAATTGGCGGGGGAGCACTACAACGGGTGGAGCCTGCGGTTTAATTGGATTCAACGCCGGGCATCTTACCA  Methanococcus
 |||||||||||||||||| ||||||  | |||| ||| |||||| ||||||  | | ||||  |||||||| || ||||||||| ||| ||||||   |   |||||| 
ACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGC.ACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCT  E. coli
    |  |||||||| ||  |||| | | |||| ||| ||||||||||  |  | | || ||  ||||||| || || || ||||| ||  |||| ||   ||||| ||| 
GTTCCGGGGGGAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCC  human
```

Figure 1–22 Genetic information conserved since the beginnings of life. A part of the gene for the smaller of the two main RNA components of the ribosome is shown. Corresponding segments of nucleotide sequence from an archaean *(Methanococcus jannaschii)*, a eubacterium *(Escherichia coli)* and a eucaryote *(Homo sapiens)* are aligned in parallel. Sites where the nucleotides are identical between species are indicated by a vertical line; the human sequence is repeated at the bottom of the alignment so that all three two-way comparisons can be seen. A dot halfway along the *E. coli* sequence denotes a site where a nucleotide has been either deleted from the eubacterial lineage in the course of evolution, or inserted in the other two lineages. Note that the sequences from these three organisms, representative of the three domains of the living world, all differ from one another to a roughly similar degree, while still retaining unmistakable similarities.



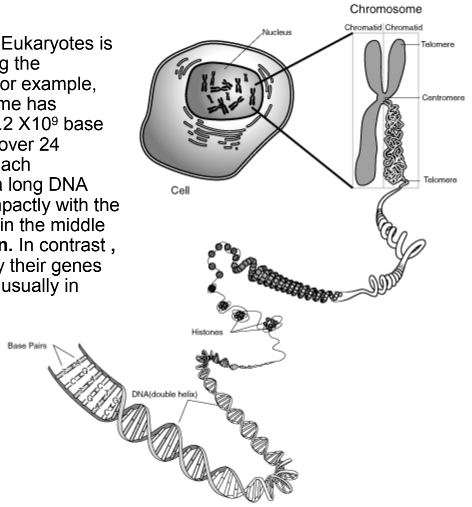
Chromosomes

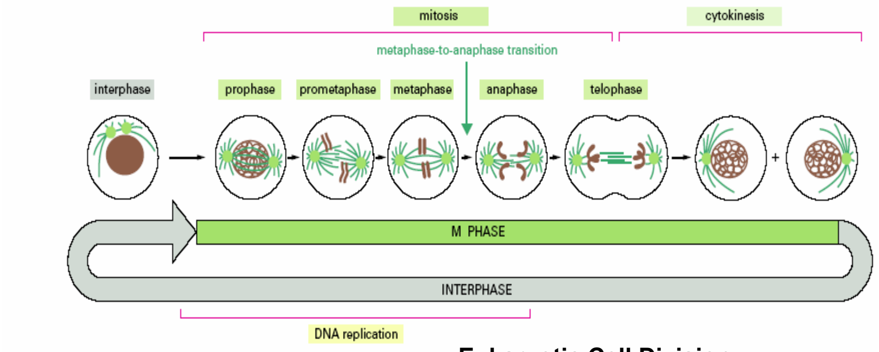
One of the threadlike "packages" of genes and other DNA in the nucleus of a cell. Different kinds of organisms have different numbers of chromosomes. Humans have 23 pairs of chromosomes, 46 in all: 44 autosomes and two sex chromosomes. Each parent contributes one chromosome to each pair, so children get half of their chromosomes from their mothers and half from their fathers

We know now that DNA in chromosome carries the hereditary Information.

The are also protein components in Chromosomes to package an enormously large string of DNA into a compact shape that can fit inside of a cell.

The DNA for the Eukaryotes is distributed among the chromosomes. For example, the human genome has approximately $3.2 X10^9$ base pairs distributed over 24 chromosomes. Each chromosome is a long DNA packed very compactly with the help of a protein in the middle called **Chromatin.** In contrast **,** the bacteria carry their genes on a single DNA usually in circular form.



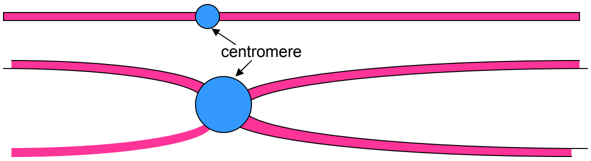


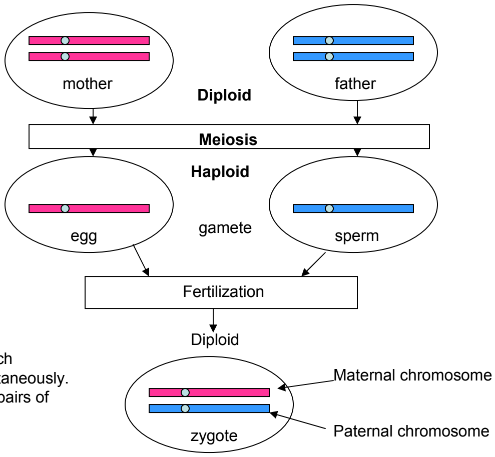
Eukaryotic Cell Division

1. The chromosomes duplicate themselves and get themselves attached to spot called *centromere*.
2. They thicken and shorten (becomes visible under microscope now).
3. The nuclear membrane dissolves and a fibrous spindle is formed, on the chromosomes liner up.
4. The centromeres divide, the spindle fibers tug the chromosome pairs apart.
5. The chromosomes gather at opposite poles, the spindle disappears.
6. The nuclear membreane re-formed, chromosomes, unwind, becomes invisible again and the two cells are formed.

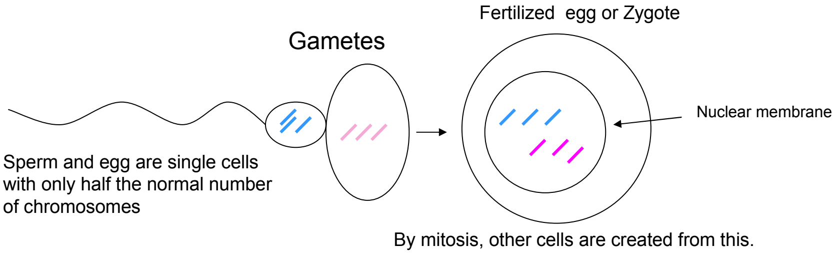
Diploid and Haploid

We explained mitotic division in the previous slide. In higher organisms, ***haploid cells***, each carrying one set of chromosomes, combine to produce ***diploid cells*** each carrying a double set of chromosomes. A chromosome is a single long double helix of DNA. At the beginning, its long and short 'arms' are joined by a **centromere**. At the end of 'metaphase', it is duplicated and condensed consisting of two identical sister double helix chain called **chromatids** joined at the centromere**.**





This diploid-haploid cycle occurs for each chromosome simultaneously. Humans have 23 pairs of Chromosomes.



Fertilized egg or Zygote

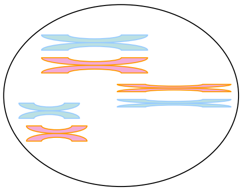
Gametes

Nuclear membrane

Sperm and egg are single cells with only half the normal number of chromosomes

By mitosis, other cells are created from this.

Each chromosome from the sperm can be matched with virtually identical one from the egg. Thus , there are two copies of each chromosome. These are called **homologs** Or homolog pairs.



Meiosis

1. As in mitosis, the chromosome doubles and thickens
2. The homologs are paired off.
3. The spindle is formed (say, in vertical direction) and the chromosome “tetrads” gather at opposite poles.
4. The pairs are separeted and reach the opposite poles.
5. The spindle vanishes and a new spindle is formed in opposite direction (say in horizontal direction.
6. The chromosomes then separate as in Mitosis.

Thus, meiosis results in four cells , each with half the number of chromosomes. Which of the homolog of each chromosome goes to which cell is completely random. This variability is significant in evolution and heredity.

Structure of DNA

DNA stands for Deoxyribonucleic acid.

A DNA is a long ***polymeric molecule***, also called a **macromolecule**.

A **polymer** is a long chain of molecules called **monomers**.

The monomer for DNA is called a **nucleotide.**

The nucleotides are components of nucleic acid.

Nucleotide is itself a very complex molecule.

Structure of DNA

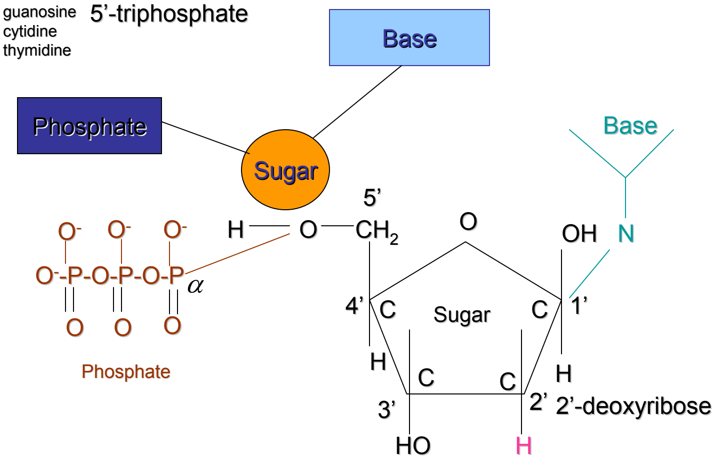
- Nucleotides are made of three basic components:
  - A sugar (deoxyribose)
  - A nitrogenous base, which is one of:
    - A – Adenine
    - G – Guanine
    - C – Cytosine
    - T – Thymine
  - A phosphate group
- DNA is a polynucleotide
- Individual nucleotides are bound by a phosphodiester bond, a covalent bond.

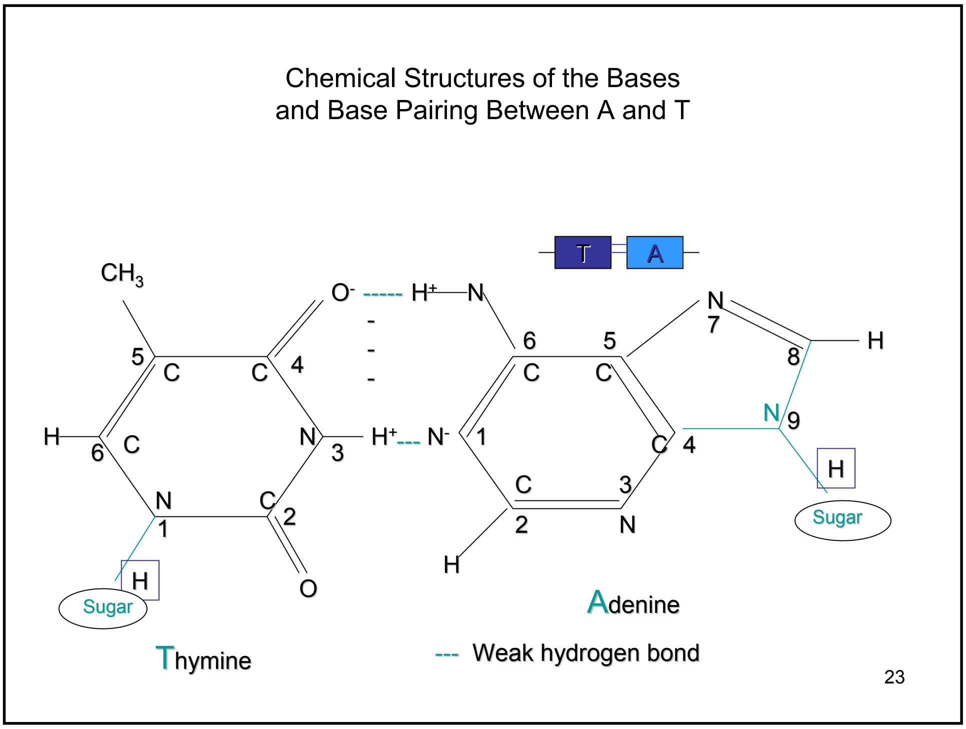
The Nitrogenous Bases

- **Purines**
  - A and G
  - double ring structures attached to 1' carbon of the sugar
  - are heavier
- **Pyrimidines**
  - C and T
  - Single ring structures attached to the 1'-carbon of the sugar
  - Lighter

Chemical Structure of a Nucleotide

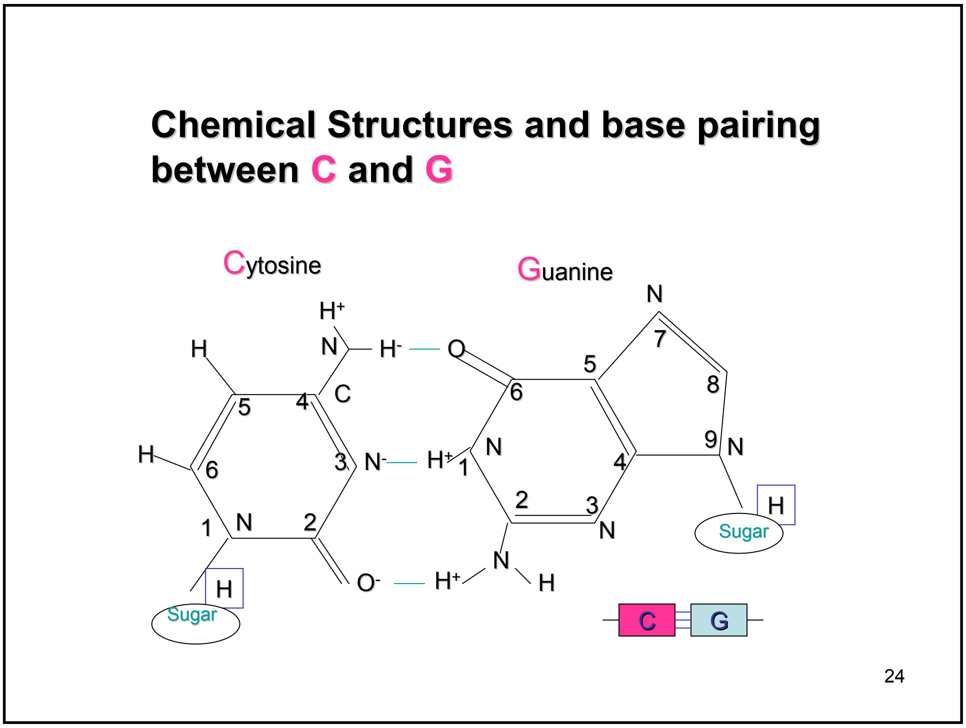
2'-deoxy adenosine
guanosine
cytidine
thymidine 5'-triphosphate





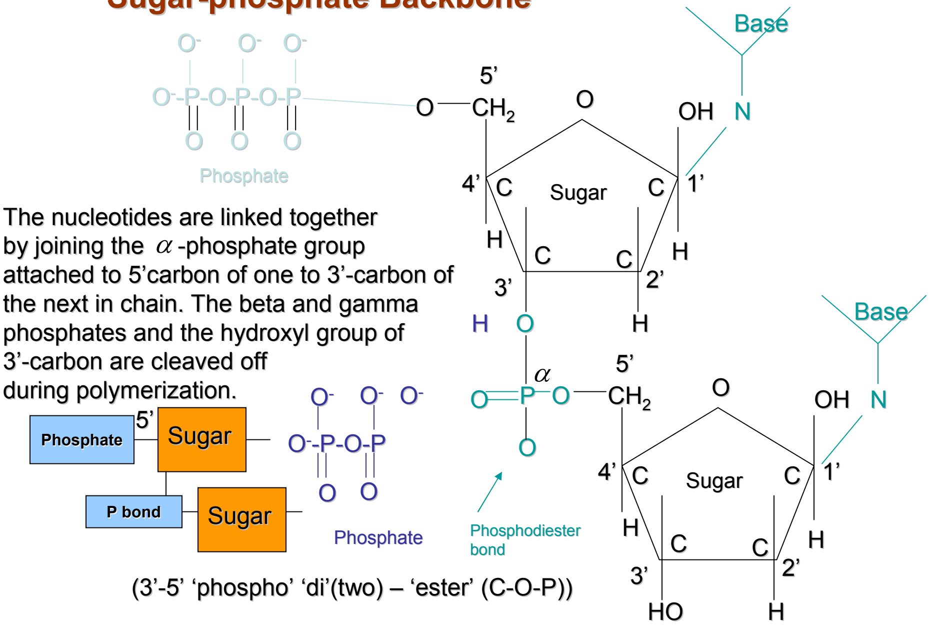
Chemical Structures of the Bases and Base Pairing Between A and T

--- Weak hydrogen bond



Chemical Structures and base pairing between C and G

Sugar-phosphate Backbone



The nucleotides are linked together by joining the α-phosphate group attached to 5'carbon of one to 3'-carbon of the next in chain. The beta and gamma phosphates and the hydroxyl group of 3'-carbon are cleaved off during polymerization.

(3'-5' 'phospho' 'di'(two) – 'ester' (C-O-P))

Orientation

The DNA or the polynucleotide has two distinct ends. The top nucleotide has the triphosphate group that does not participate in forming the phosphodiester bond. This is called the 5'-terminus. At the other end the 3'-hydroxyl group does not participate in the bond formation. This is called the 3'-terminus. These two distinct ends imply that the polynucleotides have a direction: the 5'-3' direction called the 'downstream' and the 3'-5' direction called the 'upstream'. These directions are very important as we will see later.

A String of Four Letters A, T, C, G.

Apparently, there is no limit on the number of nucleotides that can be joined together or any specific ordering of these nucleotides. Thus, the number of possible ‘strings’ that can be formed using this four letter alphabet grows exponentially.

A polynucleotide of length 10 could be any one of 4^{10}=1,048,576 possible different sequences such as

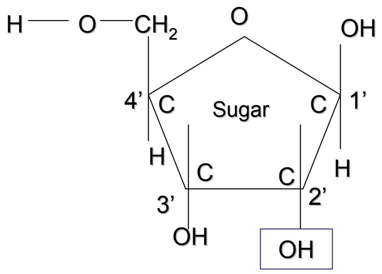
ATCGAGGTCT
GTATCCGATA

This provides potentially a very large number of variations of the genetic material if DNA length is thousands or millions.

RNA (Ribonucleic acid)

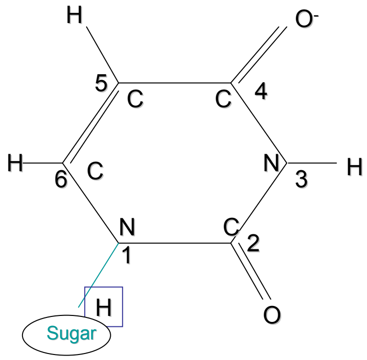
RNA, like DNA, is a polynucleotide.

1)The ‘sugar’ component in DNA (deoxyribose) is a ribose, that is, the hydroxyl group has not been ‘deoxygenated’, and



RNA

2)The base Thyamine is replaced by Uracil



Uracil

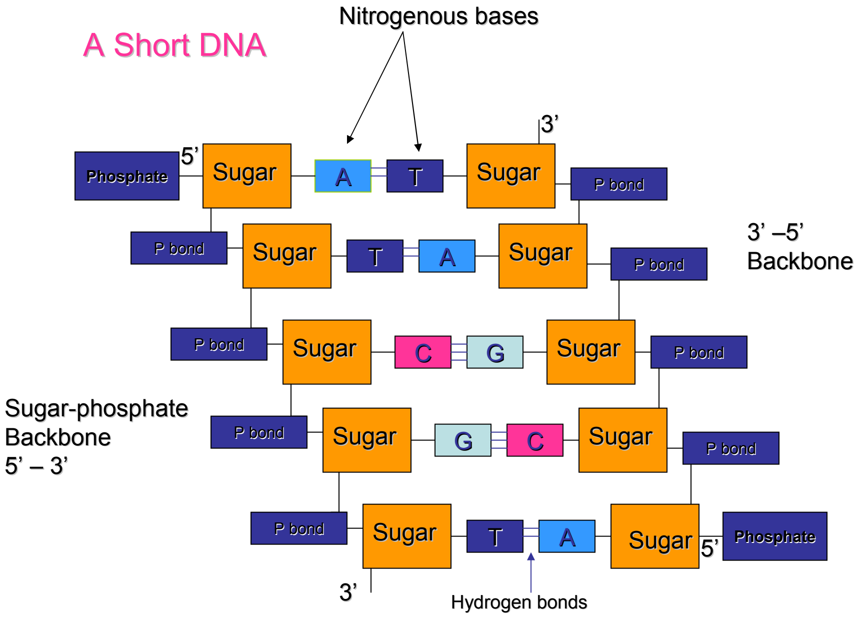
The difference between RNA and DNA

Except for the differences in ribose and uracil, the structure of RNA as a polynucleotide is the same as that of DNA. The same 3'-5' phosphodiester bond links the nucleotides in the chain and also there could be any arbitrary ordering of the four elements A,U,C and G and there is no restriction on the length of the sequence.

But , there is one very important difference. RNA in its natural form inside the cell usually exists as a single chain, but DNA exists in the form of two chains wrapped around each other in the form of a **Double Helix, a fundamental and the most important discovery in molecular biology in the early 50's.**

The 'flattened' version of the DNA structure is shown in next slide.

A Short DNA



Nitrogenous bases

Phosphate 5' Sugar A T Sugar 3' P bond

P bond Sugar T A Sugar P bond

P bond Sugar C G Sugar P bond

P bond Sugar G C Sugar P bond

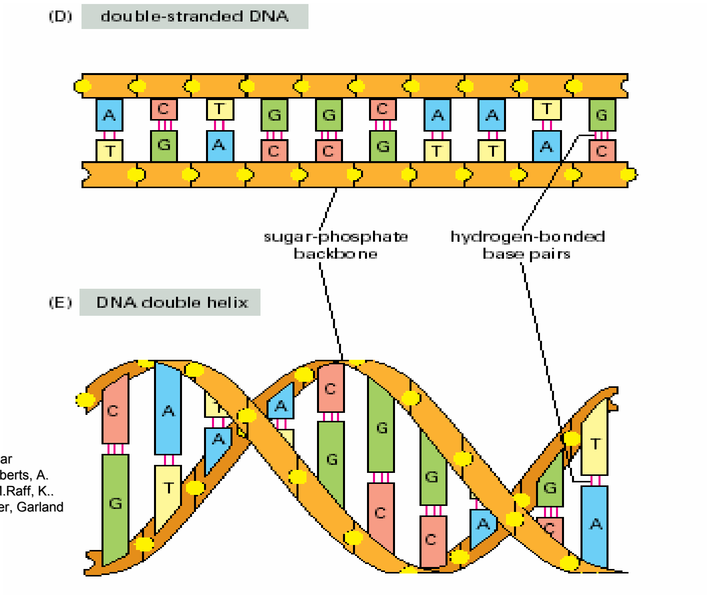
P bond Sugar T A Sugar 5' Phosphate

3'

Hydrogen bonds

3' –5'
Backbone

Sugar-phosphate
Backbone
5' – 3'



Courtesy of "Molecular Biology of Cell" B. Alberts, A. Johnson, J. Lewis, M.Raff, K.. Roberts and P. Walter, Garland Science, 4th Edition

The Story of Double helix

The discovery of double helix by James Watson and Francis Crick in 1953 was preceded by several other important discoveries in Biology:

1) Erwin Chargaff (Columbia University) found experimentally that in DNA the number of adenine residues equals the number of thyamine residues and that the number of guanine equals the number of cytosines. This is sometimes stated as A=T and C=G. But this does not mean that A+T equals C+G. In fact, the AT and CG contents of different species vary considerably. For example, the malaria parasites are very AT rich.

The Story of Double Helix

b) Rosalind Franklin of King's College, London in 1952 discovered using X-ray diffraction analysis that DNA has a helical structure. In fact, she showed that DNA exists in two forms, the A-form and the B-form. The Watson-Crick structure was the B-form of which she obtained an amazingly clear X-ray photograph.
Her research was based on earlier X-ray diffraction work pioneered by Maurice Wilkins who was her boss at the King's college. The two did not get along very well.

The photograph was shown to Watson by Wilkins without Dr. Rosalind Franklin's knowledge.

The Story of Double Helix

Francis Crick and James Watson were colleagues at Cambridge University.
Crick provided the mathematical insight and Watson,by training a physicist, was
a very aggressive young scientist who provided the power of his imagination,intuition
and modeling skill. To break the secret of DNA they were racing against time becau
Linus Pauling the towering Nobel Laureate chemist from Caltech also came close
to discovering the DNA structure (By the way, it was Rosalind Franklin who had
the courage and technical skill to point out the error in Pauling's model.)

The X-ray photograph of Rosalind Franklin provided the final clue that Watson
and Crick needed to complete their model of DNA structure. Watson and Crick
published a paper giving the correct double helix structure of DNA which earned
them Nobel prize in 1962 which was shared by Wilkins.
Rosalind was not a co-author of the paper neither was she informed of the fact
that her data was being used before the publication of their paper.
Rosalind Franklin died a sad and horrible death from cancer in 1959.

The Story of Double Helix

Further Reading:

1) "Rosalind Franklin: The Dark Lady of DNA ", Brenda Maddox, Harper Collins, 2002.

2) "The Double Helix: A Personal Account of the Discovery of the Structure of DNA" by James D. Watson and Lawrence Bragg, Atheneum, London (1968)

3) " Rosalind Franklin and DNA" by Anne Sayre, W.W.Norton, New York,1975.

4) "The Double Helix: A New Critical Edition",by G.S. Stent Weidenfeld and Nicholson, London,1981.

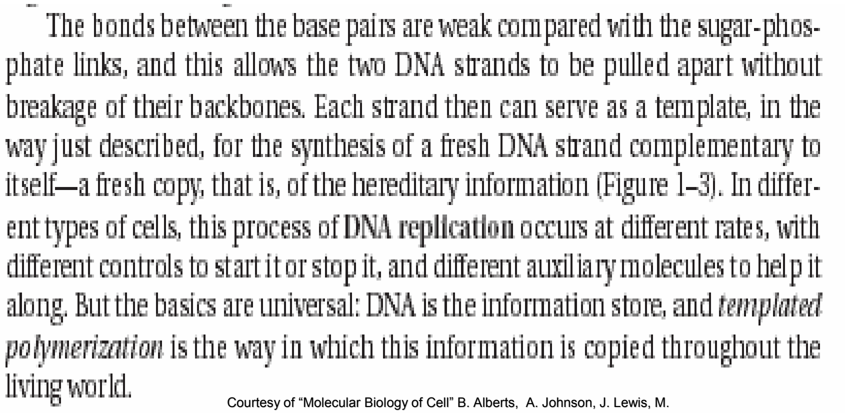
Features of the Double Helix

- Double stranded with 20 Ä in diameter.
- Bases are inside, the sugar-phosphate backbone is outside. The nitrogenous bases are stacked inside of the helix like a pile of plates.
- The two strands are held together by hydrogen bonds which can only be formed between pairs (A,T) and (C,G). These are called **complimentary base pairs**. This also explains Chargaff's ratio.
- Ten base pairs (**bp**) per turn of the helix taking 34 Ä
- The two strands are anti-parallel (5'-3' and 3'-5') and only anti-parallel polynucleotides form stable structures.
- Has two different grooves: a major grove and a minor groove.
- The double helix is right-handed

Complimentary base-pairing

- A-T
- C-G
- Other pairings not possible because:
  - Purine-Purine (A-G)
    - too big to fit in the helix
  - Pyrimidine-Pyrimidine (C-T)
    - Too small to fit in the helix
  - A-C and G-T
    - Will not align to allow hydrogen bonds to occur

Replication of DNA



The bonds between the base pairs are weak compared with the sugar-phosphate links, and this allows the two DNA strands to be pulled apart without breakage of their backbones. Each strand then can serve as a template, in the way just described, for the synthesis of a fresh DNA strand complementary to itself—a fresh copy, that is, of the hereditary information (Figure 1–3). In different types of cells, this process of **DNA replication** occurs at different rates, with different controls to start it or stop it, and different auxiliary molecules to help it along. But the basics are universal: DNA is the information store, and *templated polymerization* is the way in which this information is copied throughout the living world.

Courtesy of "Molecular Biology of Cell" B. Alberts, A. Johnson, J. Lewis, M. Raff, K.. Roberts and P. Walter, Garland Science, 4th Edition

Courtesy of "Molecular Biology of Cell" B. Alberts, A. Johnson, J. Lewis, M. Raff, K.. Roberts and P. Walter, Garland Science, 4th Edition

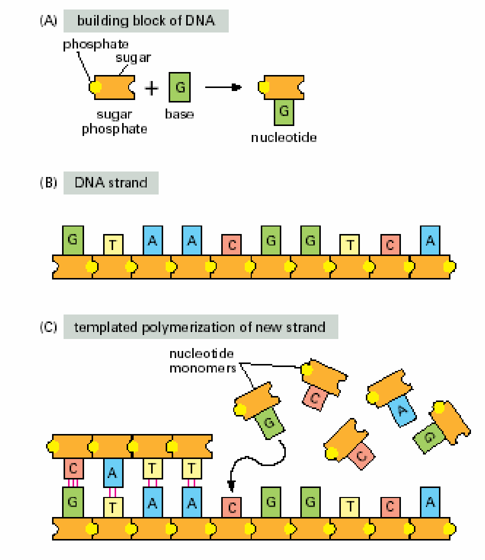
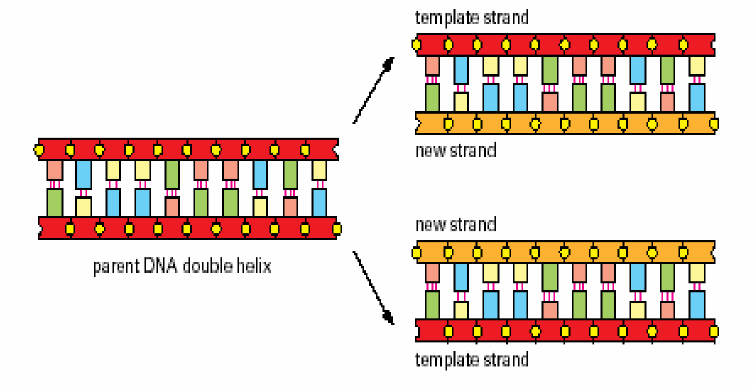


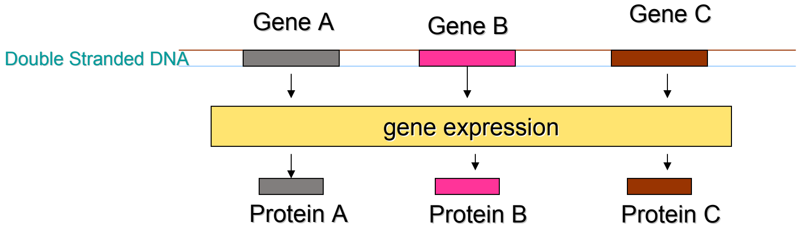
Figure 1–3 The duplication of genetic information by DNA replication. In this process, the two strands of a DNA double helix are pulled apart, and each serves as a template for synthesis of a new complementary strand.



Courtesy of "Molecular Biology of Cell" B. Alberts, A. Johnson, J. Lewis, M. Raff, K.. Roberts and P. Walter, Garland Science, 4th Edition

Genes and gene expression

A gene is a a segment of DNA molecule which codes for a single protein. It is also the functional unit of inheritance. The biological information contained in a gene acts as a set of instructions that produces a single protein via a set of intermediate processes. The entire process is called **gene expression**.



Gene A

Gene B

Gene C

Double Stranded DNA

gene expression

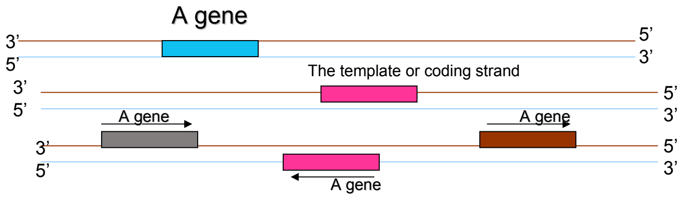
Protein A

Protein B

Protein C

DNA layout

Genes are DNA segments. The sections of the genome that contain biological information are called **exons** which are separated by vast regions of apparently useless intergenic DNA called **introns** which occupies apprximately 70% of human genome. Furthermore, the actual information is carried by only one strand of the double helix called the **template strand (**sometimes also called **coding strand).** This information is always read in the 3'-5' direction and could reside on any one of the strand.



Gene Organization

In higher organisms, genes are located in a small number of chromosomes. A chromosomes contains a long chain of a single DNA molecule or sequence (in duplicate) compactly packed around proteins. A large number of genes are Located in this one DNA strand.

Organism	Number of Chromosomes	Approx. no. of genes	Avg.no.genes per chromosomes
E.coli	1	2800	2800
Yeast	16	8750	550
Human	23	50 000	2200

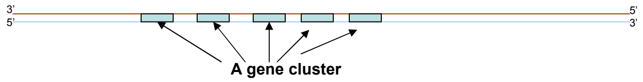
http://www.ncbi.nlm.nih.gov/genome/seq/

Gene Clusters

There are two types of gene clusters:

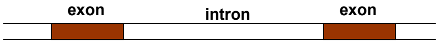
(a) Operon : occurs in bacteria. This is a cluster of genes that encodes a group of enzymes (proteins) that work collaboratively in a chemical pathway (viz. conversion of lactose absorbed by a cell Into glucose and galactose).

(b) **Multigene family**: Sometimes a single gene will occur many times in the chromosome. This is because it enhances the rate of production of a particular protein expressing the same gene in parallel. There may also be a number of similar genes that produce component structures that combine to produce a complex protein. There are also examples of gene family that are scattered over a chromosome or over more than one chromosomes.



Exons and Introns

A startling discovery was made in 1977 when several researchers found that the genes could be **split** or **discontinuous**. That is, a gene could be broken into coding regions called **exons** separated by vast amount of non-coding regions called **introns.**

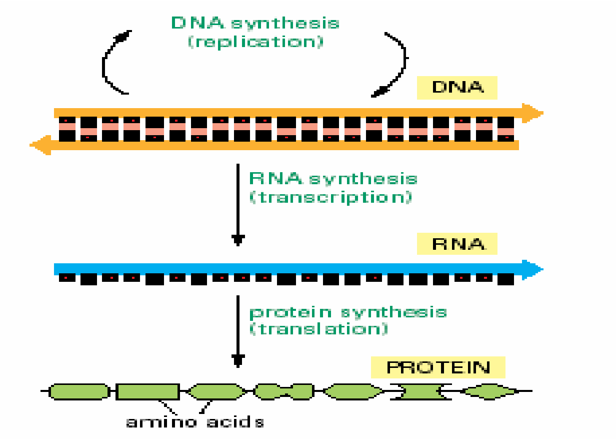


For example, the cystic fibrosis gene is 250 kb long; it is divided into 24 exons and 23 introns. The size of an exon may range from 2 to 35kb and the average length of the exon is only 227 bp. Thus about 97.6% of bp is introns, the so-called “junk DNA”. Their functions are not yet well understood.

Genome

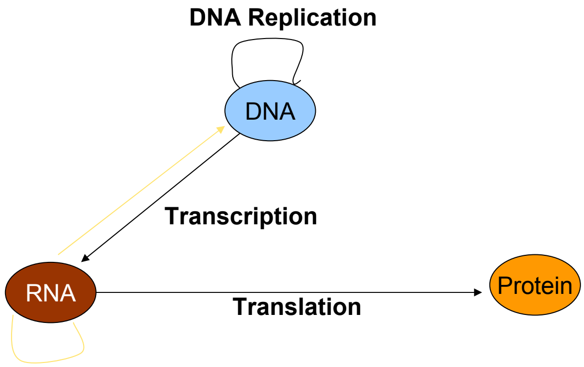
- DNA , a ‘text’ string on the four letter alphabet A, T, C and G ,spells out the biological information needed by the organism to synthesize all its proteins . Organisms differ from each other because the respective DNA molecules have different linear sequences of nucleotides which are responsible to produce different linear sequences of proteins or amino acids. There are exactly 20 different amino acids. The exact correspondence of the DNA sequence to its protein sequence is called the **genetic code** (to be discussed in detail later) and the complete sequence DNA in an organism is called its **genome**. At each cell division, the genome is copied to both its daughter cells by using the DNA duplication process explained earlier. We will now go into some more details of this process.

Figure 1–4 From DNA to protein. Genetic information is read out and put to use through a two-step process. First, in *transcription*, segments of the DNA sequence are used to guide the synthesis of molecules of RNA. Then, in *translation*, the RNA molecules are used to guide the synthesis of molecules of protein.



Transcription and Translation

The DNA in a genome does not directly produce a protein. When a cell needs a particular protein, from the very long DNA sequence in the chromosome the template strand for the protein is first copied to a corresponding RNA (by replacing the thyamines in the strand by uracils, and deoxyribose by ribose). This process is called **transcription**. (For some genes, the RNA itself might be the final product which assumes a 3-dimensional structure after **folding.** Such RNAs play structural and catalytic roles in the cell.) The information in RNA , **now called messenger RNA or mRNA**, is then used to synthesize a **polypeptide or a linear sequence of amino acids** which folds into a 3-dimensional **protein** structure**.** This process of gene expression is universal from bacteria to humans and has been termed the **central dogma** of molecular biology.



Crick's Central Dogma

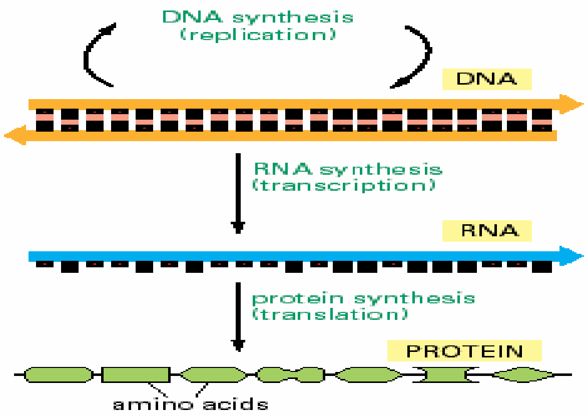


Figure 1–4 From DNA to protein. Genetic information is read out and put to use through a two-step process. First, in *transcription*, segments of the DNA sequence are used to guide the synthesis of molecules of RNA. Then, in *translation*, the RNA molecules are used to guide the synthesis of molecules of protein.

Protein Functions

Proteins are multi-functional elements of all living organism. They could be:

(a) Structural proteins: bones, cartilage, tendons.
(b) Contractile proteins: muscles
(c) Enzymes: catalyses other bio-chemical functions
(d) Regulatory proteins: control and regulate bio-chemical reactions
(e) Protective proteins : immunoglobins and antibodies
(f) Storage proteins: ovalbumin, ferritin etc.

Homologs

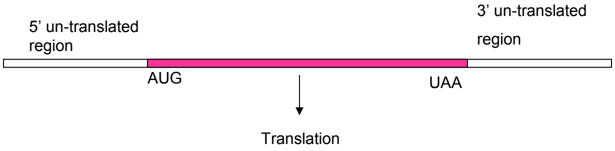
For higher organisms with sex discrimination, each cell contains two copies of each chromosome, one inherited from mother and the other inherited from father. This pair is called **homologous chromosomes or homologs.** Homologs are DNA sequences with a high degree of similarity. The only non-homologous chromosomes are the sex chromosomes in males, where father contributes a Y Chromosome and the mother contributes a X chromosome. Each human cells contain a total of 46 chromosomes- 22 pairs common to both male and female, X and Y in males and two X's in female.

Gene expression

We will start with the simple situation of transcription of a single gene to a RNA. During transcription, the RNA transcript is built up in a step by step fashion using the DNA template as a guide. The template is read in 3'-5' direction and the RNA synthesis takes place almost like the DNA duplication using complimentary strand except that the 'complement' of A is U, Uracil. But the formation of the phosphodiester bond follows the same chemical principles. The enzyme that catalyses the process is called **RNA polymerase**. The polymerase is a complex macromolecule consisting of about 7000 molecules for prokaryotes like *E.Coli.* For eukaroyotes, this polymerase is much more complex.

A schematic representation of the transcription process is shown next.

Anatomy of RNA



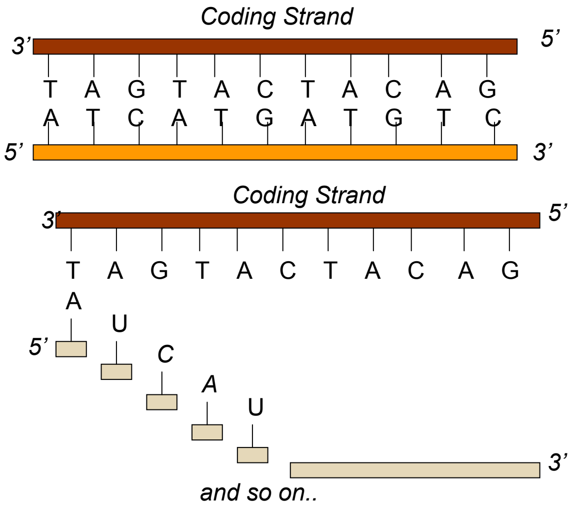
5’ un-translated region

3’ un-translated region

AUG

UAA

Translation



Coding Strand

3’ 5’

T A G T A C T A C A G

A T C A T G A T G T C

5’ 3’

Coding Strand

3’ 5’

T A G T A C T A C A G

5’ A U C A U 3’

and so on..

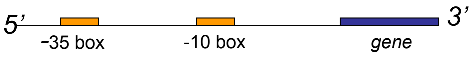
The final one-stranded RNA : AUCAUGAUGUC

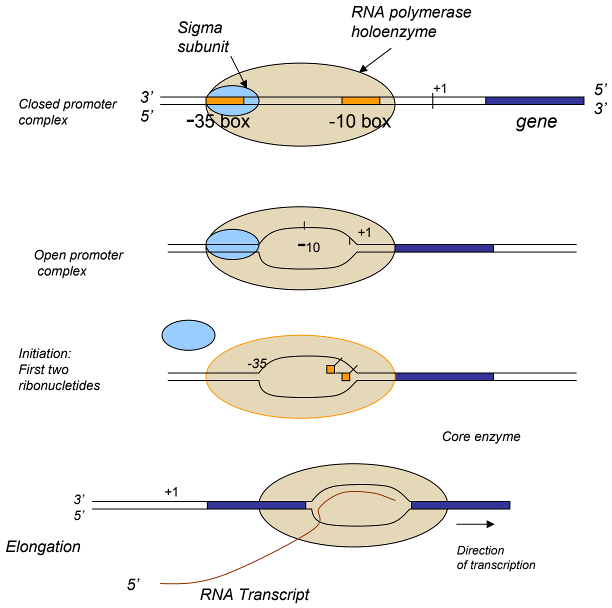
Transcription for E.Coli

The transcription process in reality even for a simple E.Coli is much more complex than what we have described. The process is divided into three phases: **initiation, elongation and termination.**

Initiation: The RNA polymerase initiates the operation and it must transcribe not any arbitrary part of DNA but only the gene. For this the polymerase first 'bind' to a location **upstream** of the gene. This site is called a **promoter** sequence. The promoter is a short DNA sequence which can be bind to the polymerase. In E.Coli, the promoter sequence consists of two distinct sequences at a distance -10 and -35 upstream from the position at which transcription starts. The actual sequences may vary from gene to gene but they are related to the following two **consensus sequences** both located in the non-template strand**:**

-35 box 5'-TTGACA-3'
-10 box 5'-TATAAT-3'





Steps of RNA Transcription

The sigma subunit within the polymerase recognizes the promoter sequence and a **closed promoter complex** is formed. The enzyme covers about 60 bp of the double helix. In the next phase, the double helix starts 'melting' at -10 box unwinding the DNA into single strands in the region under the core enzyme. The -10 box consistsof entirely AT pairs which have only two hydrogen bonds for each bp. This makes it makes it easier to melting to take place compared to the situation with CG base pairs which have 3 hydrogen bonds. This configuration is called **open promoter complex,** the sigma subunit ejects out of the holoenzyme converting it to a core enzyme. At the same time, the first two ribonucleotides are sealed in the template strand at positions +1 and +2 with a phosphodiester bond.

In the next **elongation step**, the polymerase moves downstream with relative ease, unzipping the DNA molecule and attaching new ribonucleotides to the 3' end of the growing RNA. At the same time, the DNA behind it rebounds back to its double helix structure. The open promoter is

like a bubble that propagate to 3' direction always maintaining its size between 12 to 17 DNA. Also, the rate of propagation is not constant, it may slow down, pause, reaccelerate or even go backwards destroying the ribonucleotides. The RNA itself is synthesized in the 5'-3' direction. The length of the actual transcript is longer than the length of the gene because the +1 position is about 20 to 600 nucleotide upstream from the beginning of the gene. This part of the RNA transcript is called a **leader segment**. Similarly, the transcription extends beyond the gene creating a similar **trailer segment**.

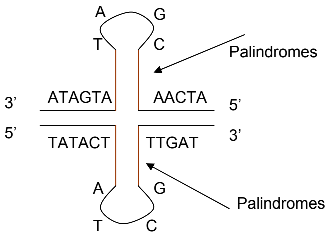
The **termination** of RNA transcription is signalled by the presence of a **complementary palindrome.** (A palindrome reads the same sequence in both forward and backward direction viz ATAGCGATA. The complementary palindrome is ATAGCCTAT.). This means that within each strand of the DNA and within its RNA transcript, base pairing might occur. Le's take an example,

3'---ATATGA CGTCCGA TAGC TCGGACG AACTA ---5'
5'---TATACT GCAGGCT ATCG AGCCTGC TTGAT---3'

The RNA resulting from this DNA will be

5'---UAUACU GCAGGCU AUCG AGCCUGC UUGAU---3'

which forms a stem-loop or 'hair-pin 'structure' as



Cruciform structure of DNA

```
5'---UAUACU      UUGAU---3'
          G  C
          C  G
          A  U
          G  C
          G  C
          C  G
          U  A
        A      G
          U C
```

Stem-loop
Structure

The formation of the stem-loop structure has proved to be a difficult subject for the biologists to study. Some stem-loop terminators are followed by by a run of 5 to 10 A's in the DNA template and results in number of U's in the RNA transcript just before termination. This structure sends a controlling signal that detaches the RNA from the template. Alternately, the presence of a very large protein call **Rho** which attaches to the growing RNA might stop the motion of the Polymmerase and terminate the RNA transcript.

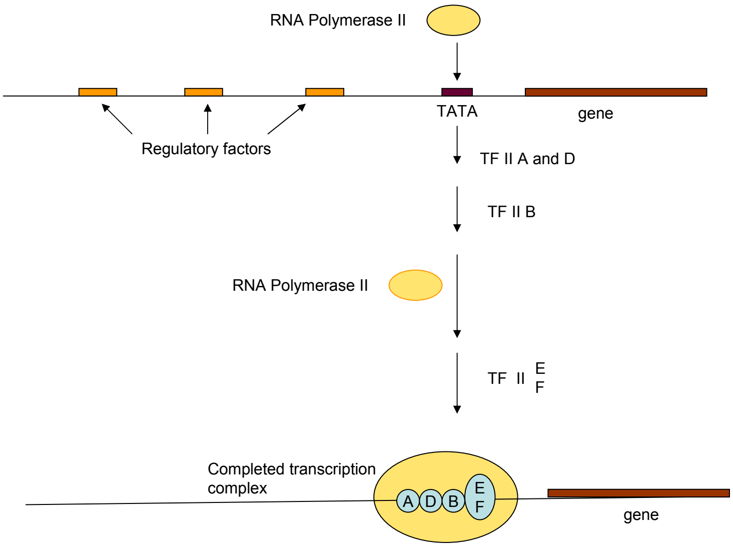
At the end of the termination, the RNA polymerase falls off the DNA, recombines with its sigma units, the 'melted' DNA part recoils back to its double helix configuration.

Transcription in Eukaroytes

The transcription in Eukaryotes is similar to that in E.Coli but is much more complex. The RNA polymerase has an **attachment site** rather than a promoter sequence plus other promoter sequences distributed over several hundred base pairs all upstream from the genes. These promoters regulate the gene expression by turning on or off the transcription process. Understanding these regulatory processes is by itself a whole new research field.

The attachment site is referred to as **-25 TATA box** (5'-TATAAAT-3') and the RNA polymerase called **RNA Polymerase II.** The attachment is helped by a set of proteins called **transcription factors (TF II A, TF II B , D, E and F**) which ultimately makes the transcription complex ready to start the synthesis process of RNA. The next slide gives a schematic representation.

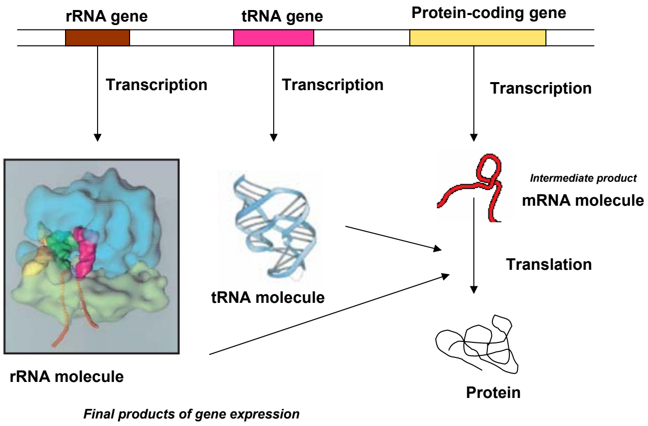
The exact details of the termination of the transcription process is not very well understood. The termination trail seem to be longer (about 1000 to 2000 bp downstream the gene. The exact termination point is still a matter of research.



mRNA, tRNA and rRNA

The ***messenger RNA or mRNA*** acts as an intermediary between the DNA and protein synthesis and they are short-lived and are produced whenever the organism needs to produce new proteins. They are not the end products of gene expression. The other two RNAs, the ***transfer RNA or tRNA*** and the **ribosomal RNA or rRNA** , in contrast, are final end products and they are long-lived and referred to as *stable* RNAs. Both play crucial roles in the gene expression.

The three major types of RNA molecules produced by transcription



rRNA gene

tRNA gene

Protein-coding gene

Transcription

Transcription

Transcription

Intermediate product

mRNA molecule

Translation

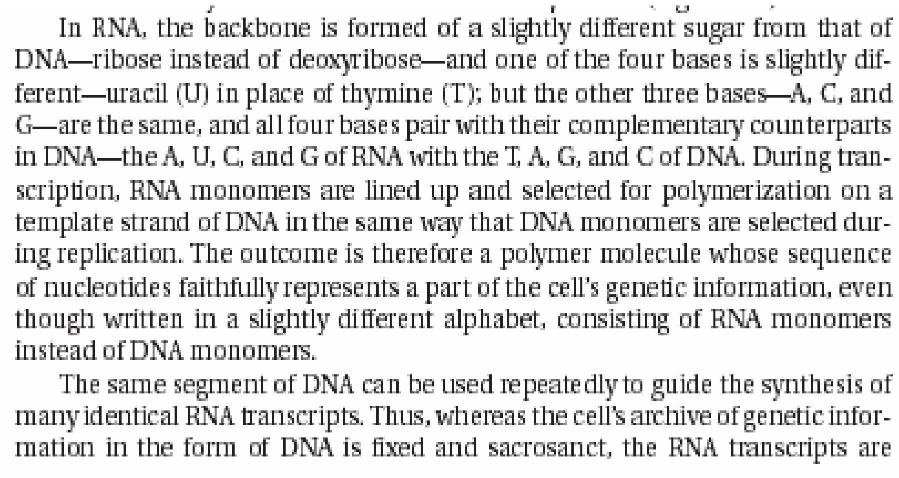
tRNA molecule

rRNA molecule

Protein

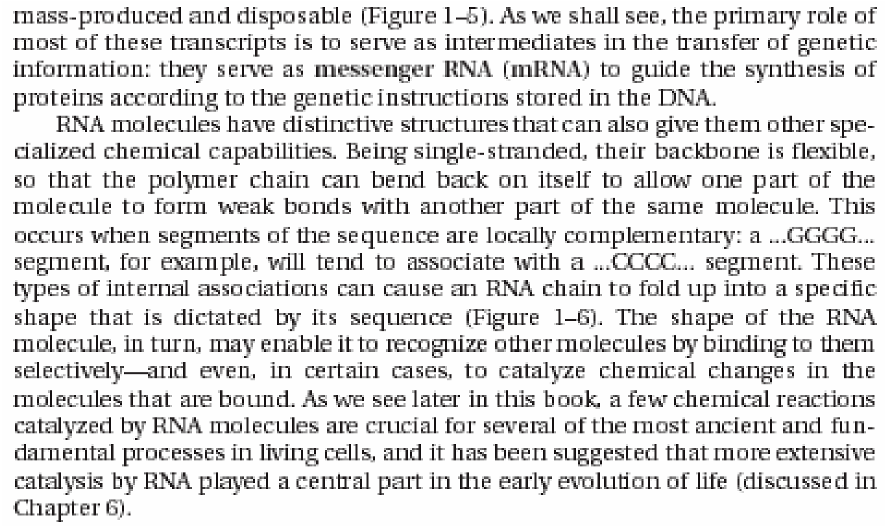
Final products of gene expression

We will explain the functions of these RNA by directly quoting from the seminal text book "Cell". We will add some detail information derived from both of our reference text.



In RNA, the backbone is formed of a slightly different sugar from that of DNA—ribose instead of deoxyribose—and one of the four bases is slightly different—uracil (U) in place of thymine (T); but the other three bases—A, C, and G—are the same, and all four bases pair with their complementary counterparts in DNA—the A, U, C, and G of RNA with the T, A, G, and C of DNA. During transcription, RNA monomers are lined up and selected for polymerization on a template strand of DNA in the same way that DNA monomers are selected during replication. The outcome is therefore a polymer molecule whose sequence of nucleotides faithfully represents a part of the cell's genetic information, even though written in a slightly different alphabet, consisting of RNA monomers instead of DNA monomers.

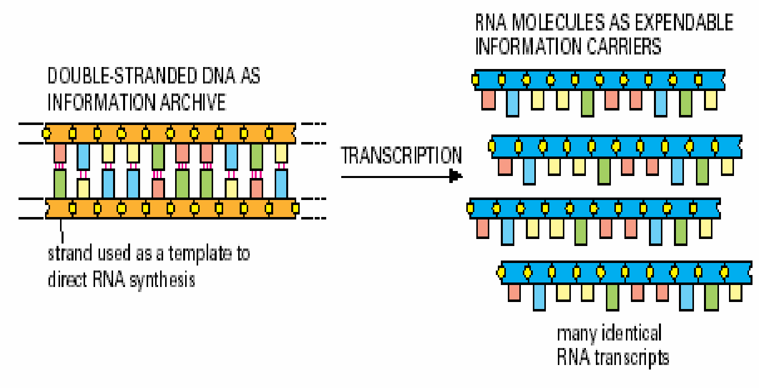
The same segment of DNA can be used repeatedly to guide the synthesis of many identical RNA transcripts. Thus, whereas the cell's archive of genetic information in the form of DNA is fixed and sacrosanct, the RNA transcripts are



mass-produced and disposable (Figure 1–5). As we shall see, the primary role of most of these transcripts is to serve as intermediates in the transfer of genetic information: they serve as **messenger RNA (mRNA)** to guide the synthesis of proteins according to the genetic instructions stored in the DNA.

RNA molecules have distinctive structures that can also give them other specialized chemical capabilities. Being single-stranded, their backbone is flexible, so that the polymer chain can bend back on itself to allow one part of the molecule to form weak bonds with another part of the same molecule. This occurs when segments of the sequence are locally complementary: a ...GGGG... segment, for example, will tend to associate with a ...CCCC... segment. These types of internal associations can cause an RNA chain to fold up into a specific shape that is dictated by its sequence (Figure 1–6). The shape of the RNA molecule, in turn, may enable it to recognize other molecules by binding to them selectively—and even, in certain cases, to catalyze chemical changes in the molecules that are bound. As we see later in this book, a few chemical reactions catalyzed by RNA molecules are crucial for several of the most ancient and fundamental processes in living cells, and it has been suggested that more extensive catalysis by RNA played a central part in the early evolution of life (discussed in Chapter 6).

Figure 1–5 How genetic information is broadcast for use inside the cell. Each cell contains a fixed set of DNA molecules—its archive of genetic information. A given segment of this DNA serves to guide the synthesis of many identical RNA transcripts, which serve as working copies of the information stored in the archive. Many different sets of RNA molecules can be made by transcribing selected parts of a long DNA sequence, allowing each cell to use its information store differently.


RNA MOLECULES AS EXPENDABLE INFORMATION CARRIERS
DOUBLE-STRANDED DNA AS INFORMATION ARCHIVE
TRANSCRIPTION
strand used as a template to direct RNA synthesis
many identical RNA transcripts

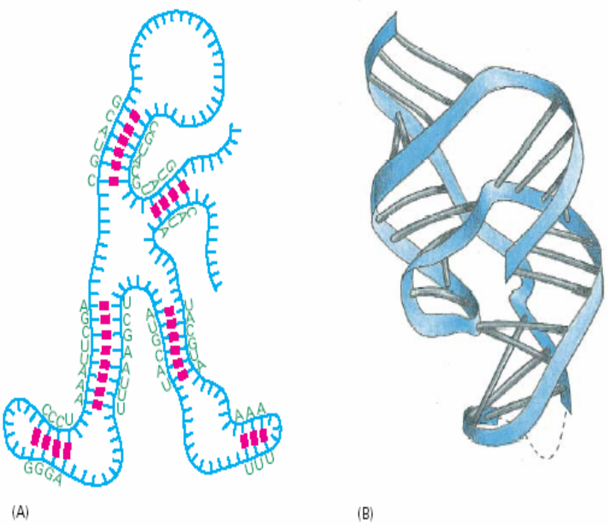


Figure 1–6 The conformation of an RNA molecule. (A) Nucleotide pairing between different regions of the same RNA polymer chain causes the molecule to adopt a distinctive shape. (B) The three-dimensional structure of an actual RNA molecule from hepatitis delta virus that catalyzes RNA strand cleavage. (B, based on A.R. Ferré D'Amaré, K. Zhou, and J.A. Doudna, *Nature* 395:567–574, 1998. © Macmillan Magazines Ltd.)

All Cells Use Proteins as Catalysts

Protein molecules, like DNA and RNA molecules, are long unbranched polymer chains, formed by the stringing together of monomeric building blocks drawn from a standard repertoire that is the same for all living cells. Like DNA and RNA, they carry information in the form of a linear sequence of symbols, in the same way as a human message written in an alphabetic script. There are many different protein molecules in each cell, and—leaving out the water—they form most of the cell's mass.

The monomers of protein, the **amino acids**, are quite different from those of DNA and RNA, and there are 20 types, instead of 4. Each amino acid is built around the same core structure through which it can be linked in a standard way to any other amino acid in the set; attached to this core is a side group that gives each amino acid a distinctive chemical character. Each of the protein molecules, or **polypeptides**, created by joining amino acids in a particular sequence folds into a precise three-dimensional form with reactive sites on its surface (Figure 1–7A). These amino acid polymers thereby bind with high specificity to other molecules and act as **enzymes** to catalyze reactions in which covalent bonds are made and broken. In this way they direct the vast majority of chemical processes in the cell (Figure 1–7B). Proteins have a host of other functions as well—maintaining structures, generating movements, sensing signals, and so on—each protein molecule performing a specific function according to its own genetically specified sequence of amino acids. Proteins, above all, are the molecules that put the cell's genetic information into action.

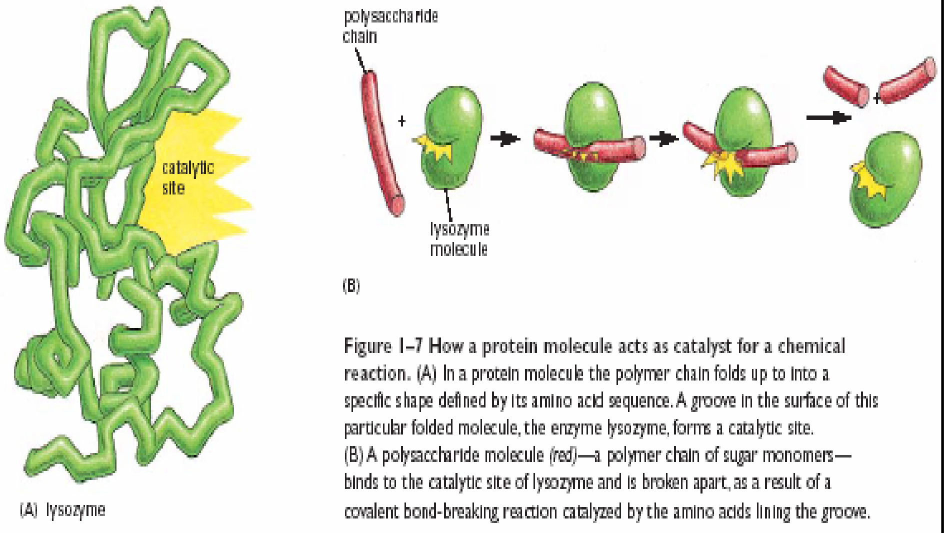
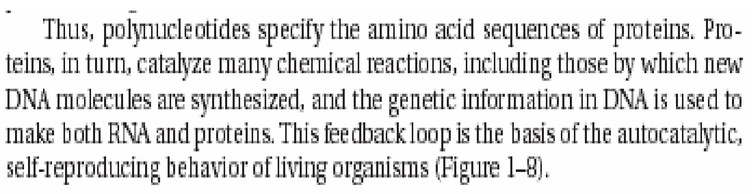


Figure 1–7 How a protein molecule acts as catalyst for a chemical reaction. (A) In a protein molecule the polymer chain folds up to into a specific shape defined by its amino acid sequence. A groove in the surface of this particular folded molecule, the enzyme lysozyme, forms a catalytic site. (B) A polysaccharide molecule *(red)*—a polymer chain of sugar monomers—binds to the catalytic site of lysozyme and is broken apart, as a result of a covalent bond-breaking reaction catalyzed by the amino acids lining the groove.



Thus, polynucleotides specify the amino acid sequences of proteins. Proteins, in turn, catalyze many chemical reactions, including those by which new DNA molecules are synthesized, and the genetic information in DNA is used to make both RNA and proteins. This feedback loop is the basis of the autocatalytic, self-reproducing behavior of living organisms (Figure 1–8).

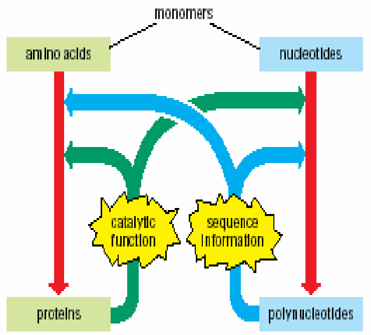
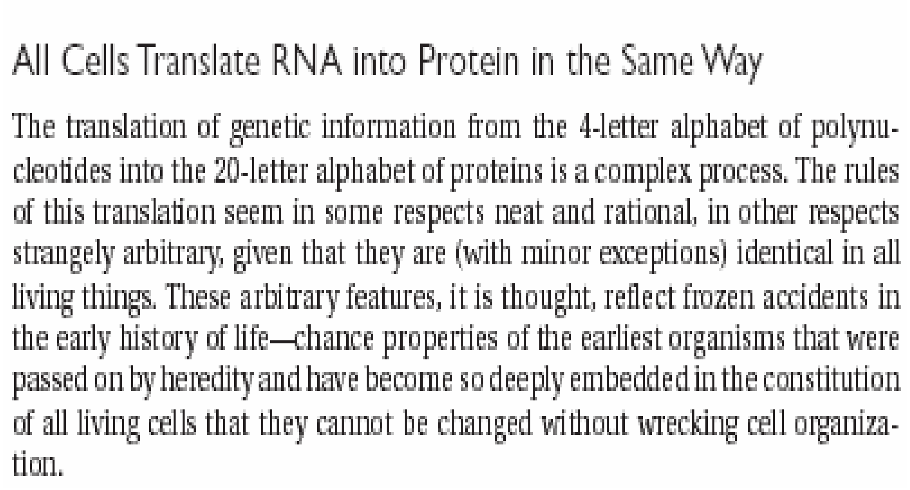


Figure 1–8 Life as an autocatalytic process. Polynucleotides (nucleotide polymers) and proteins (amino acid polymers) provide the sequence information and the catalytic functions that serve—through a complex set of chemical reactions—to bring about the synthesis of more polynucleotides and proteins of the same types.



All Cells Translate RNA into Protein in the Same Way

The translation of genetic information from the 4-letter alphabet of polynucleotides into the 20-letter alphabet of proteins is a complex process. The rules of this translation seem in some respects neat and rational, in other respects strangely arbitrary, given that they are (with minor exceptions) identical in all living things. These arbitrary features, it is thought, reflect frozen accidents in the early history of life—chance properties of the earliest organisms that were passed on by heredity and have become so deeply embedded in the constitution of all living cells that they cannot be changed without wrecking cell organization.

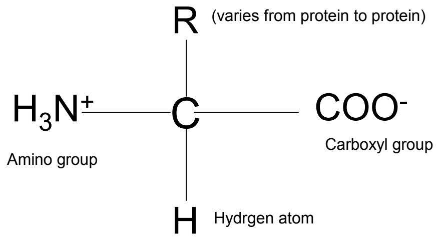
Proteins

Proteins are polymers, also called polypeptides consisting of a sequence of amino acids. There are twenty amino acids that are found in proteins.

Hydrophobic Group			**Hydrophilic Group**		
A	**Alanine**	**ala**	**R**	**Arginine**	**arg**
C	**Cysteine**	**cys**	**N**	**Asparagine**	**asn**
G	**Glycine**	**gly**	**D**	**Aspartic acid**	**asp**
I	**Isoleucine**	**ile**	**Q**	**Glutamine**	**gln**
L	**Leucine**	**leu**	**E**	**Glutamic acid**	**glu**
M	**Methionine**	**met**	**H**	**Histidine**	**his**
F	**Phenylalanine**	**phe**	**K**	**Lysine**	**lys**
P	**proline**	**pro**	**S**	**Serine**	**ser**
T	**Trypyophan**	**trp**	**T**	**Threonine**	**thr**
Y	**Tyrosine**	**tyr**			
V	**Valine**	**val**			

Chemical Structure of Proteins

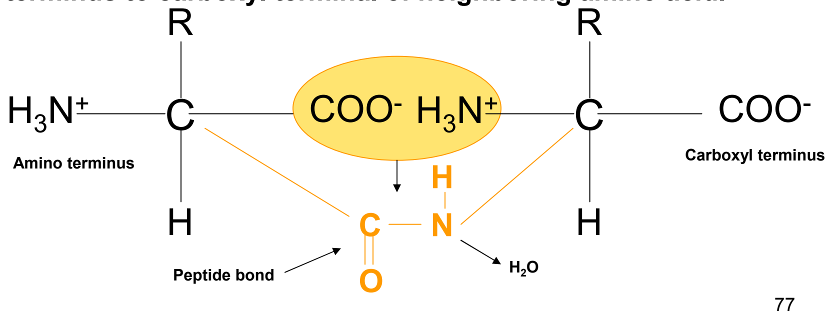
Each protein has a general structure. It consists of
A central carbon atom and four groups attached to it.



The R groups vary considerably in complexity. For gly it is a single hydrogen atom. For tyr and phe , it is a complex side chain. For details, see the text references.

Peptide Bonds

The DNA polymer is formed by phophodiester bonds. For amino acid, the amino acids are bonded by a peptide bond which is The reason proteins are also called polypeptide. The bonding structure is shown below.The polypeptide grows from its Nitrogen terminus to carboxyl terminal of neighboring amino acid.



Protein Structures and Functions

The protein molecule assumes a complex 3 dimensional structures. The primary structure is the amino acid sequenc. There are two kinds of secondary structures called α-helix and β -sheets, both stabilized at the terminus by the hydrogen bond of the amino and carboxyl groups.There are also tertiary and quaternary structures.
A new field has emerged for the study of structures and functions of proteins called *proteomics* . The proteins play important regulatory roles in gene expression, DNA-binding and a host of other functions. We will have occasions to discuss some of these later. Thus, we have an autocatalytic system and an information flow system with feedback as shown in Fig.1-8 earlier

The Genetic Code

We have stated earlier: one gene, one protein. But, how a gene defines a protein uniquely? It took several years since the discovery of double helix by famous biologists (all Nobel Laureates: Severo Ochoa, Marshall Nirenberg and Gobind Khorana, Crick and Brenner) to decipher what is called the *genetic code*. The first observation was the order of nucleotides in the gene directly determine the order of amino acid in the polypeptide (protein).

ATCTGTAACGGATAT ← DNA sequence in gene

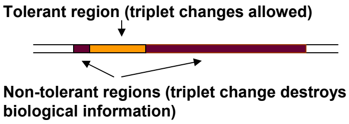
AUCUGUAACGGAUAU ← mRNA

I C N G Y ← polypeptide sequence

The second important discovery was about the size of the code word. Since DNA alphabet has 4 symbols (A,C,T,G) [although U rather than T appears in mRNA, the code is stated in terms of DNA alphabet], two Symbols can code at most 4^2=16 polypeptide, so we need at least three letters since there are 20 proteins. Three it is – from elementary computer science principle – giving an excess number of 4^3=64 combinations! Nature has used all 64 combinations incorporating redundancy and robustness!! But, establishing this by biological experiments was an extremely challenging task.

Reading Frame

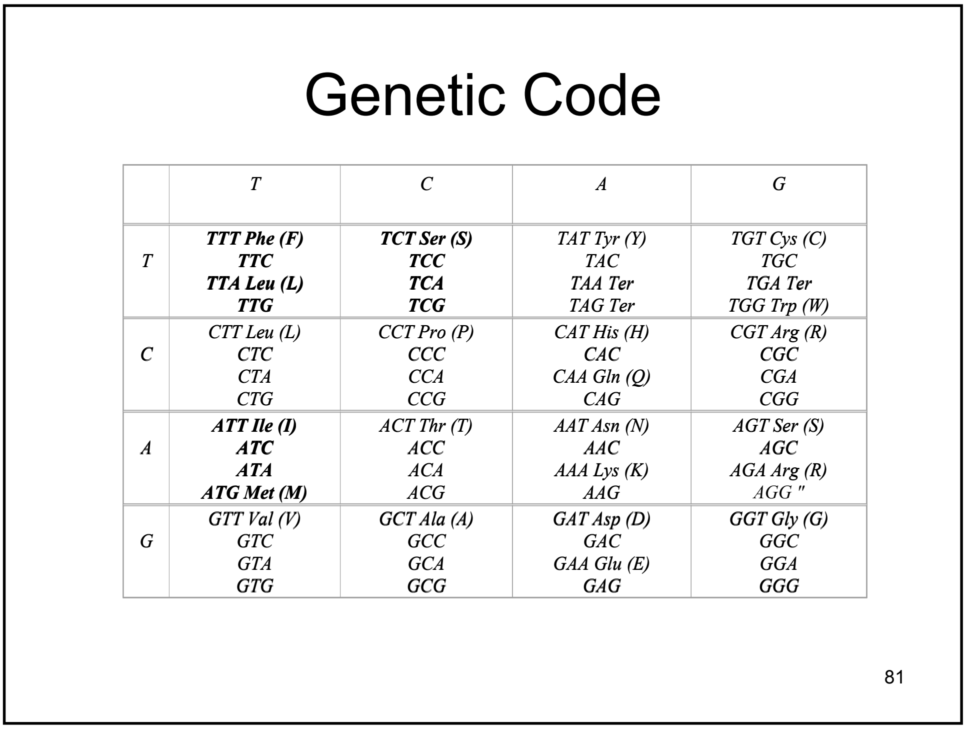
The linearity of gene and protein was established by a simple experiment. (You may come up with this experiment if you think a bit)
But, the experiment to proving the triplet nature of the code took some detailed Biology work. First there are these acridine dyes that can cause deletion or insertion of just one base pair in a double stranded DNA. It was also discovered that there are some proteins in which a segment of amino acids can be changed without altering its function. This portion of the protein is called ***tolerant** regions*.
If regions of the gene corresponding to this tolerant region, is changed one at time what effect will have? If the code is a triplet and we delete or insert one bp in the gene, then all the amino acids downstream will be changed yielding a non-functional protein. If we delete or insert two bp in the gene, it will also have the same effect. But if we insert or delete three bp, then the correct **reading frame** in non-tolerant region will be restored and the protein will be functional again. Crick and Brenner performed an elegant experiment based on this princple and established the triplet nature of gene coding. Nirenberg, Holley and Govind Khorana actually later deciphered the genetic code table and shared Nobel prize in 1968.



Tolerant region (triplet changes allowed)

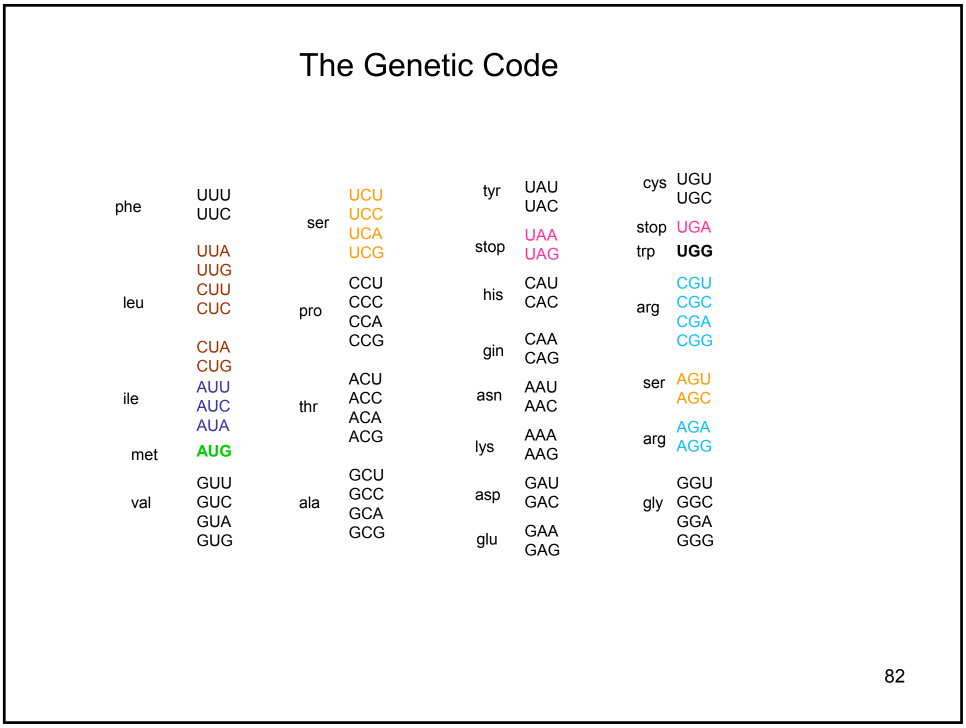
Non-tolerant regions (triplet change destroys biological information)

CAN YOU SEE THE WAY ITS DUN
CAN YOX USE ETH EWA YIT SDU N
CAN YOX UZS EET HEW AYI TSD UN
CAN YOX UZS QEE THE WAY ITS DUN



Genetic Code

	T	C	A	G
T	**TTT Phe (F)** **TTC** **TTA Leu (L)** **TTG**	**TCT Ser (S)** **TCC** **TCA** **TCG**	TAT Tyr (Y) TAC TAA Ter TAG Ter	TGT Cys (C) TGC TGA Ter TGG Trp (W)
C	CTT Leu (L) CTC CTA CTG	CCT Pro (P) CCC CCA CCG	CAT His (H) CAC CAA Gln (Q) CAG	CGT Arg (R) CGC CGA CGG
A	**ATT Ile (I)** **ATC** **ATA** **ATG Met (M)**	ACT Thr (T) ACC ACA ACG	AAT Asn (N) AAC AAA Lys (K) AAG	AGT Ser (S) AGC AGA Arg (R) AGG "
G	GTT Val (V) GTC GTA GTG	GCT Ala (A) GCC GCA GCG	GAT Asp (D) GAC GAA Glu (E) GAG	GGT Gly (G) GGC GGA GGG



The Genetic Code

phe	UUU UUC	ser	UCU UCC UCA UCG	tyr	UAU UAC	cys	UGU UGC
leu	UUA UUG CUU CUC CUA CUG	pro	CCU CCC CCA CCG	stop	UAA UAG	stop	UGA
ile	AUU AUC AUA	thr	ACU ACC ACA ACG	his	CAU CAC	trp	**UGG**
met	**AUG**	ala	GCU GCC GCA GCG	gin	CAA CAG	arg	CGU CGC CGA CGG
val	GUU GUC GUA GUG			asn	AAU AAC	ser	AGU AGC
				lys	AAA AAG	arg	AGA AGG
				asp	GAU GAC	gly	GGU GGC GGA GGG
				glu	GAA GAG		

Code Features

All possible 64 triplets have a meaning.

The codes are not unique for a particular amino acid except for *met (AUG)* and *trp (UGG)*.

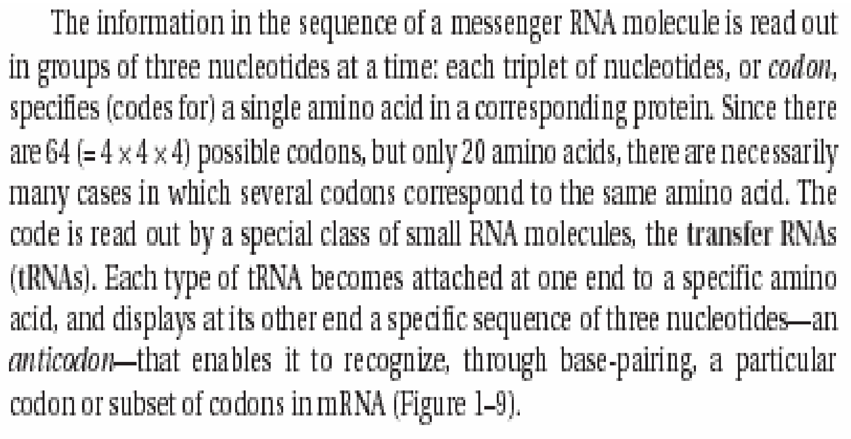
If an amino acid has multiple codes, its first two letters are the same.

The codes UAA, UAG and UGA denote '***stop***' and do not represent any protein.

The code AUG appears at the beginning of every protein and it is therefore recognized as an **initiation code** but also has the significance that all proteins are synthesized with a methionine at the beginning (but later may be discarded within the cell). Methionine may not be always an initiation code and may also appear in the middle of a protein.

The genetic code is not universal in the sense that it applies only to nuclear genes. The mitochondrial genes use a slightly different code and also expressed differently using different ribosome and tRNAs.

Translation



The information in the sequence of a messenger RNA molecule is read out in groups of three nucleotides at a time: each triplet of nucleotides, or *codon*, specifies (codes for) a single amino acid in a corresponding protein. Since there are 64 (= 4 × 4 × 4) possible codons, but only 20 amino acids, there are necessarily many cases in which several codons correspond to the same amino acid. The code is read out by a special class of small RNA molecules, the **transfer RNAs** (tRNAs). Each type of tRNA becomes attached at one end to a specific amino acid, and displays at its other end a specific sequence of three nucleotides—an *anticodon*—that enables it to recognize, through base-pairing, a particular codon or subset of codons in mRNA (Figure 1–9).

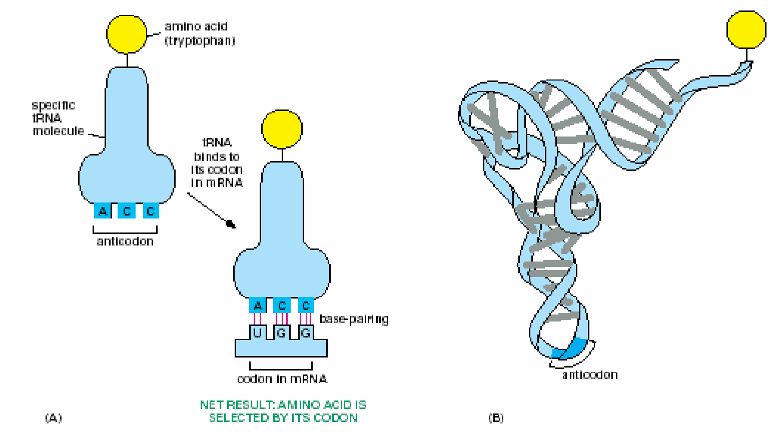
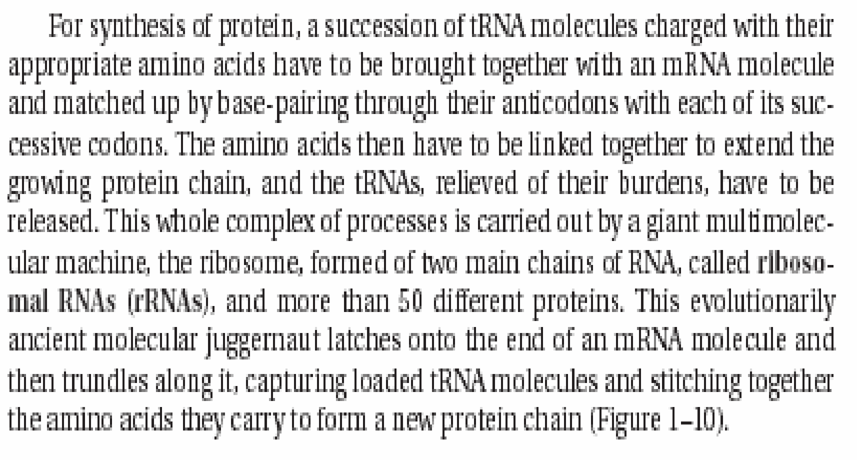
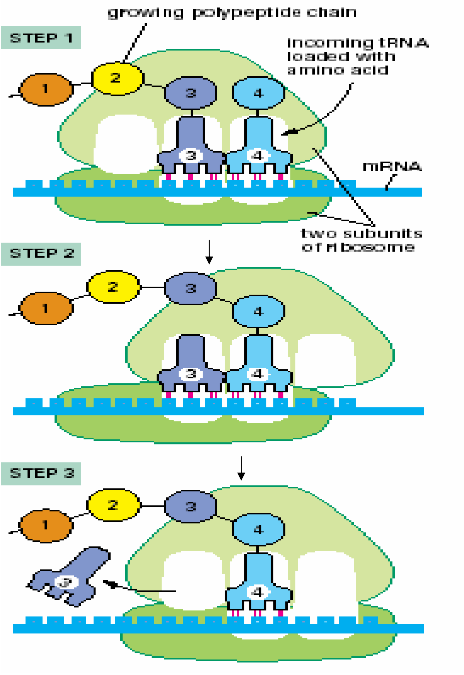


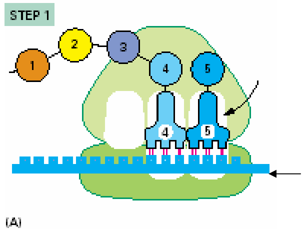
Figure 1–9 Transfer RNA. (A) A tRNA molecule specific for the amino acid tryptophan. One end of the tRNA molecule has tryptophan attached to it, while the other end displays the triplet nucleotide sequence CCA (its anticodon), which recognizes the tryptophan codon in messenger RNA molecules. (B) The three-dimensional structure of the tryptophan tRNA molecule. Note that the codon and the anticodon in (A) are in antiparallel orientations, like the two strands in a DNA double helix (see Figure 1–2), so that the sequence of the anticodon in the tRNA is read from right to left, while that of the codon in the mRNA is read from left to right.



For synthesis of protein, a succession of tRNA molecules charged with their appropriate amino acids have to be brought together with an mRNA molecule and matched up by base-pairing through their anticodons with each of its successive codons. The amino acids then have to be linked together to extend the growing protein chain, and the tRNAs, relieved of their burdens, have to be released. This whole complex of processes is carried out by a giant multimolecular machine, the ribosome, formed of two main chains of RNA, called **ribosomal RNAs (rRNAs)**, and more than 50 different proteins. This evolutionarily ancient molecular juggernaut latches onto the end of an mRNA molecule and then trundles along it, capturing loaded tRNA molecules and stitching together the amino acids they carry to form a new protein chain (Figure 1–10).

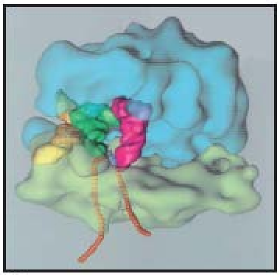
Figure 1-10 A ribosome at work. (A) The diagram shows how a ribosome moves along an mRNA molecule, capturing tRNA molecules that match the codons in the mRNA and using them to join amino acids into a protein chain. The mRNA specifies the sequence of amino acids. (B) The

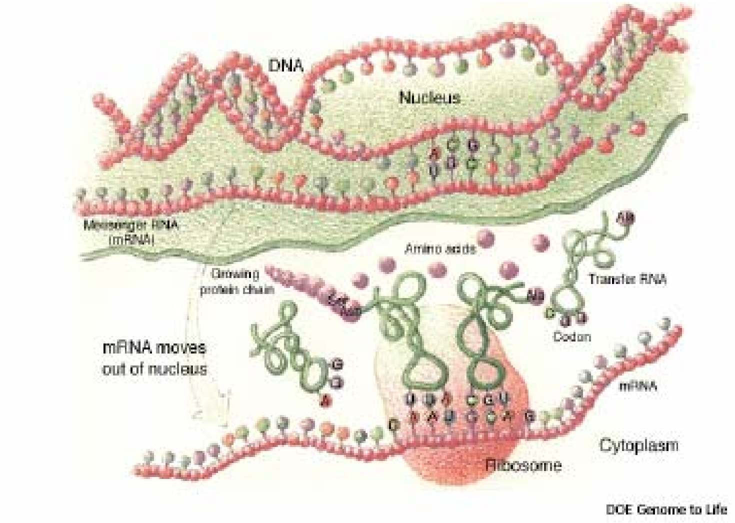




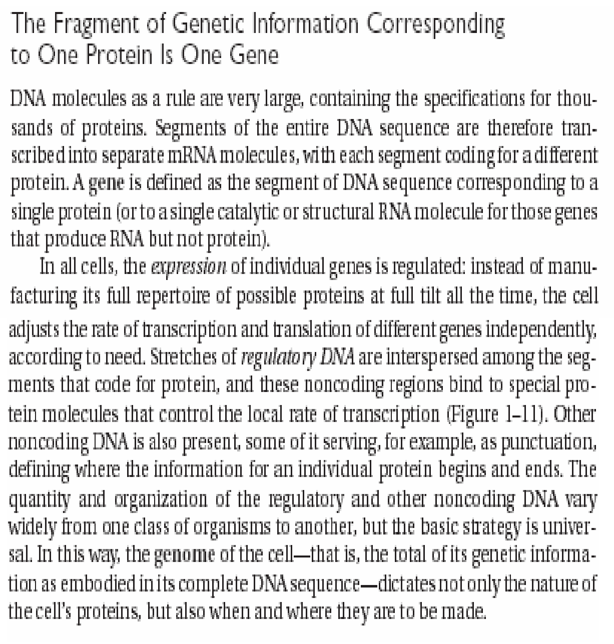
1.10 (B) The

three-dimensional structure of a bacterial ribosome (pale green and blue), moving along an mRNA molecule (orange beads), with three tRNA molecules (yellow, green, and pink) at different stages in their process of capture and release. The ribosome is a giant assembly of more than 50 individual protein and RNA molecules. (B, courtesy of Joachim Frank, Yanhong Li, and Rajendra Agarwal.)


(B)



DOE Genome to Life



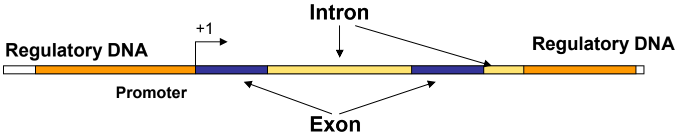
The Fragment of Genetic Information Corresponding to One Protein Is One Gene

DNA molecules as a rule are very large, containing the specifications for thousands of proteins. Segments of the entire DNA sequence are therefore transcribed into separate mRNA molecules, with each segment coding for a different protein. A **gene** is defined as the segment of DNA sequence corresponding to a single protein (or to a single catalytic or structural RNA molecule for those genes that produce RNA but not protein).

In all cells, the *expression* of individual genes is regulated: instead of manufacturing its full repertoire of possible proteins at full tilt all the time, the cell adjusts the rate of transcription and translation of different genes independently, according to need. Stretches of *regulatory DNA* are interspersed among the segments that code for protein, and these noncoding regions bind to special protein molecules that control the local rate of transcription (Figure 1–11). Other noncoding DNA is also present, some of it serving, for example, as punctuation, defining where the information for an individual protein begins and ends. The quantity and organization of the regulatory and other noncoding DNA vary widely from one class of organisms to another, but the basic strategy is universal. In this way, the **genome** of the cell—that is, the total of its genetic information as embodied in its complete DNA sequence—dictates not only the nature of the cell's proteins, but also when and where they are to be made.

Gene Regulation

The genome in a single cell contains a staggering amount of information. The DNA sequence in the Genome not only includes the recipes for proteins but also the conditions under which the recipe is invoked. DNA also encodes the RNA molecules (mRNA, tRNA) needed for protein synthesis. DNA molecules are not directly involved in synthesizing sugars and lipids but produce enzymes that can catalyze their production(metabolic action). Thus, a gene is really a set of substrings of DNA template strand that are responsible for the entire process of synthesis of a protein, called gene expression.

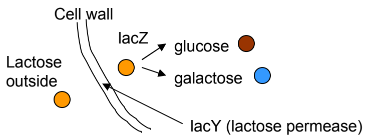


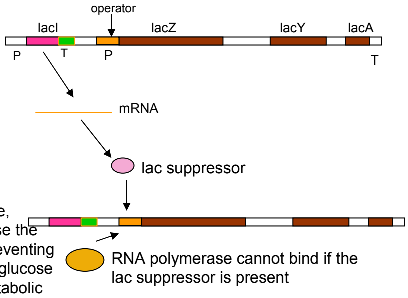
Transcription of a gene can can be regulated or enhanced by DNA binding proteins both up- and downstream of the gene**.** Not all genes for a given cell are active at all times. Some of the genes have to be continuously active viz. the genes that produce ribosomal RNA or tRNA or genes that code for enzymes such as RNA polymerase or those that are involved in metabolic actions. These are called **housekeeping genes**. But, most genes act as need arises due to environmental stimulus and specific function of the cell in the organism. This means there must be some **control mechanism** to turn on or off the genes. Even in bacteria such **regulatory genes have** been found**.** In response to different kinds of sugar in the environment, different genes become expressed. Yeast creates different enzymes by expressing specific genes for specific kinds of sugars. Plants express photosynthetic gents in response to sun light. For multi-cellular organisms, certain groups of cells respond to hormones produced by a different groups of cells. Specialized cells always express a subset of genes required for its specific functions. For higher organism, the shape and size of body parts are controlled by growth hormones which have to be stopped at some point by turning off the genes that produced such hormones. How does it happen? The process is only partially understood.

Gene Regulation

Gene regulation occurs mostly at the transcription level although for the housekeeping genes regulation takes place at the levels of RNA in order to take into account the required rate of production to maintain a stable state.

As an example of gene regulation in prokaryotes , biologists have studied in detail a cluster of 3 genes (jointly called an operon) that control utilization of lactose(milk sugar) in E.coli.
lacZ: **encodes an enzyme that splits lactose into glucose and galactose.**
lacY : **encodes for a protein that helps pump lactose from outside environment to inside of the cell**
lacA : encodes for an enzyme that degrades carbon compounds to extract energy.
There is also an additional gene *lacI* with its own promoter and *t*erminator which acts as a repressor which prevents *RNA polymerase to attach to the DNA sequence stopping transcription of the operon.*





lacI acts as a regulatory gene. If there is no lactose, transcription of lac operaon does not occur because the Lac suppressor attaches to the operator region preventing RNA polymerase to bind. If there is also sufficient glucose within the cell, a second regulatory gene called catabolic activator protein (CAP) just upstream the operator, prevents RNA polymerase to bind inactivating the operon.

Eukaroytic Genes

1. Only 1.1 percent of the genome codes for protreins and RNA molecules
2. There are many repeats
3. The genes are not contiguous; exons are separated by long segments of introns. For example, the two genes BRCA-1 and BRCA-2, responsible for breast cancer in chromosome 17 has about 98% introns. Their lengths are about 100,000 and 200,000 bp but code for only 1863 and 3418 amino acids, respectively.
4. The gene regulation is much more complex, a multiple number of them exist both upstream and downstream, as far as 50000 bp away.

The Eukaryotic proteins that regulate transcription by binding to regulatory sites are called transcription factor.

Metallothionein Gene

Metallothionein is a protein that protects cells when exposed to carcinogenic heavy metal like cadmium.