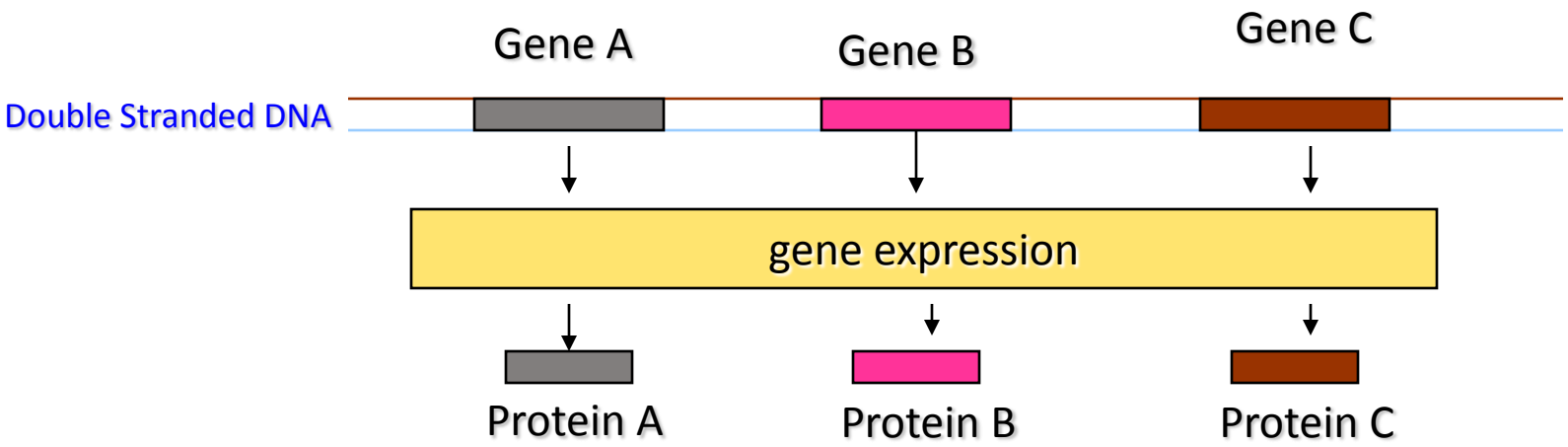
Molecular Biology
Part III

What carries information between DNA to Proteins

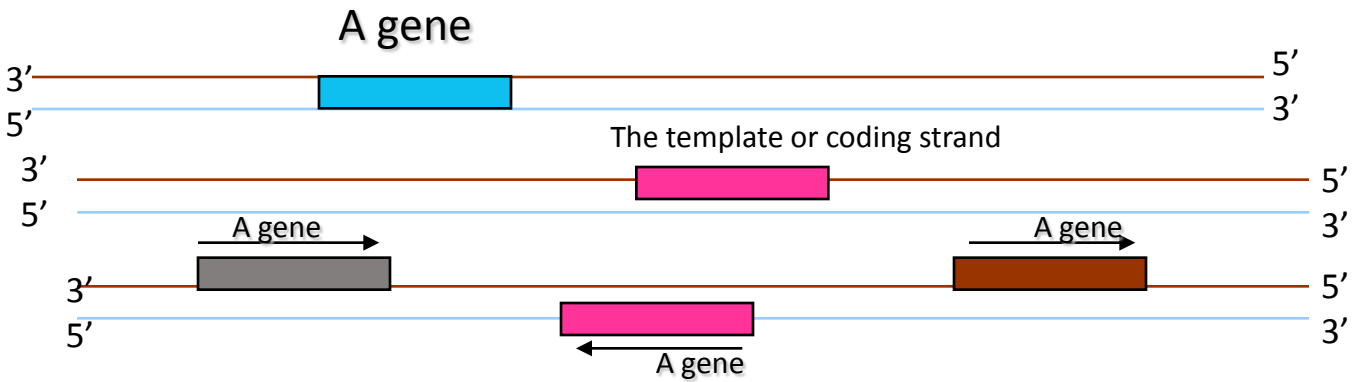
Genes and gene expression

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**.



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.



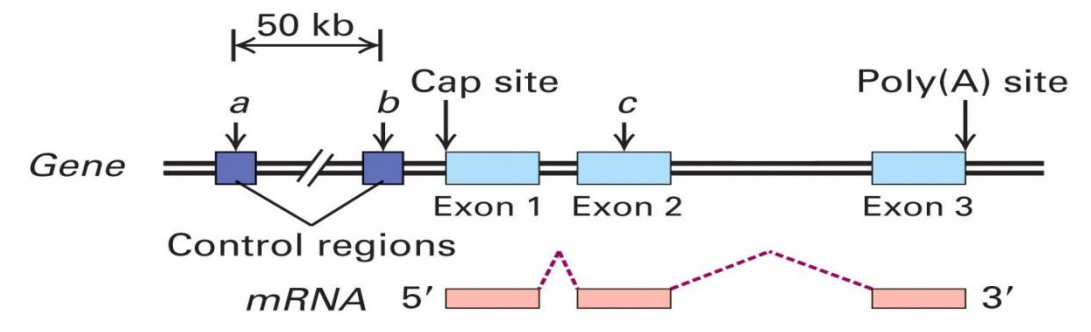
Gene Organization

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

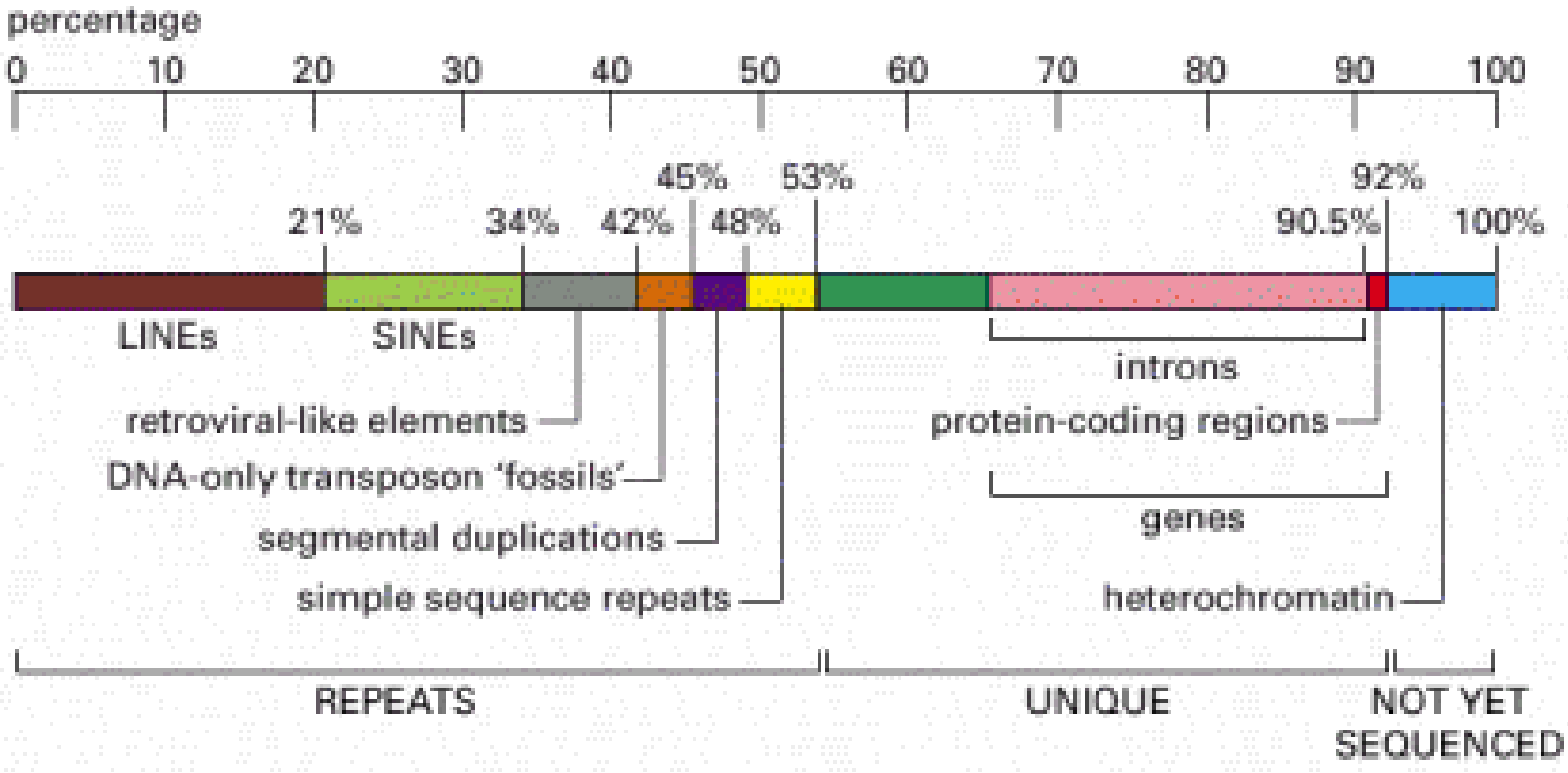
http://www.ncbi.nlm.nih.gov/genome/seq/

Definition of a Gene



- Regulatory regions: up to 50 kb upstream of +1 site
- Exons: protein coding and untranslated regions (UTR)
  1 to 178 exons per gene (mean 8.8)
  8 bp to 17 kb per exon (mean 145 bp)
- Introns: splice acceptor and donor sites, junk DNA
  average 1 kb – 50 kb per intron
- Gene size: Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.

Human genome sequence

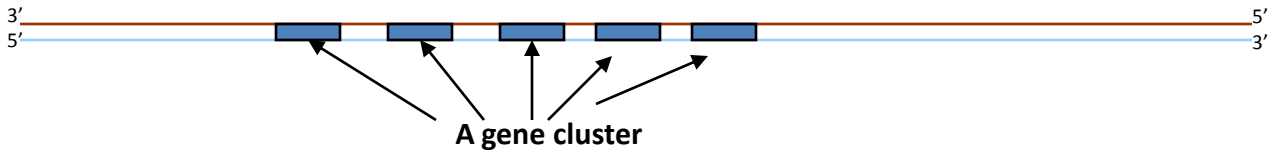


Gene Clusters

There are two types of gene clusters:

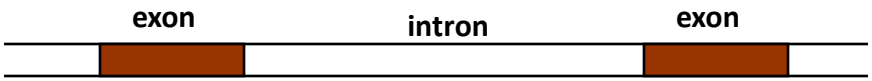
(a) Operon : occurs in bacteria. This is a cluster of genes that encodes a group of enzymes (proteins) that work collaboratively in a chemical pathway (viz. conversion of lactose absorbed by a cell Into glucose and galactose).

(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.**

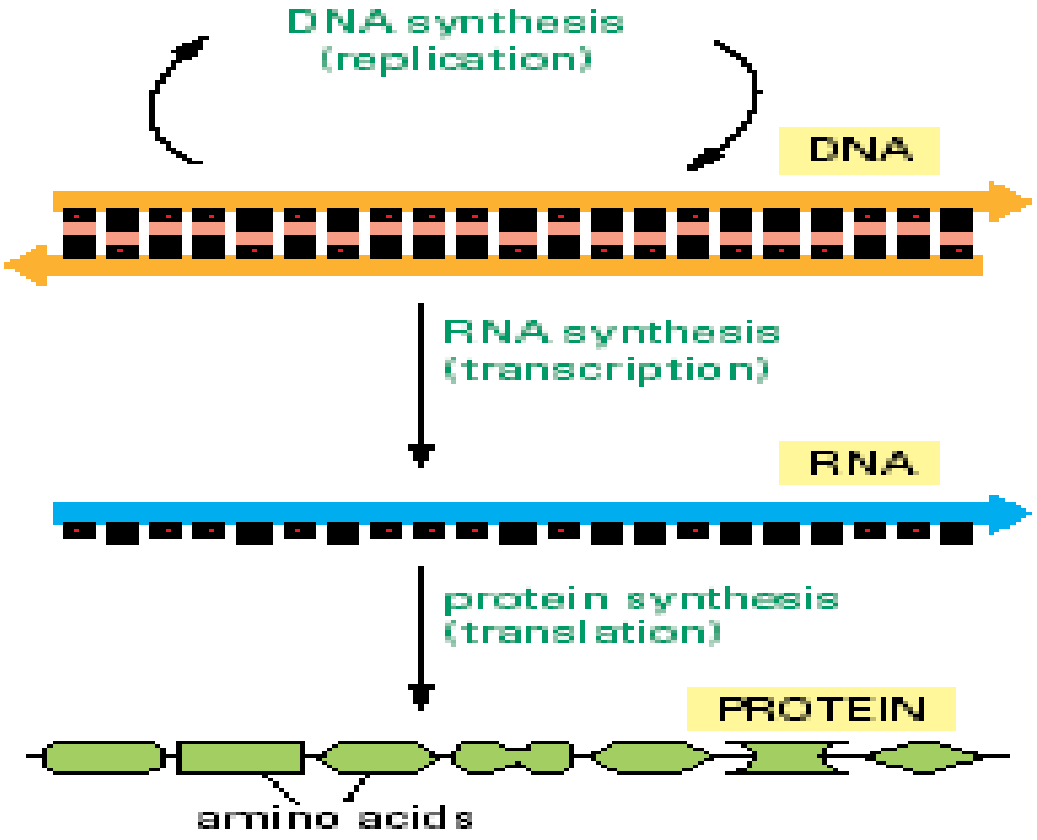


For example, the cystic fibrosis gene is 250 kb long; it is divided into 24 exons and 23 introns. The size of an exon may range from 2 to 35kb and the average length of the exon is only 227 bp. Thus about 97.6% of bp is introns, the so-called “junk DNA”. Their functions are not yet well understood.

Genome

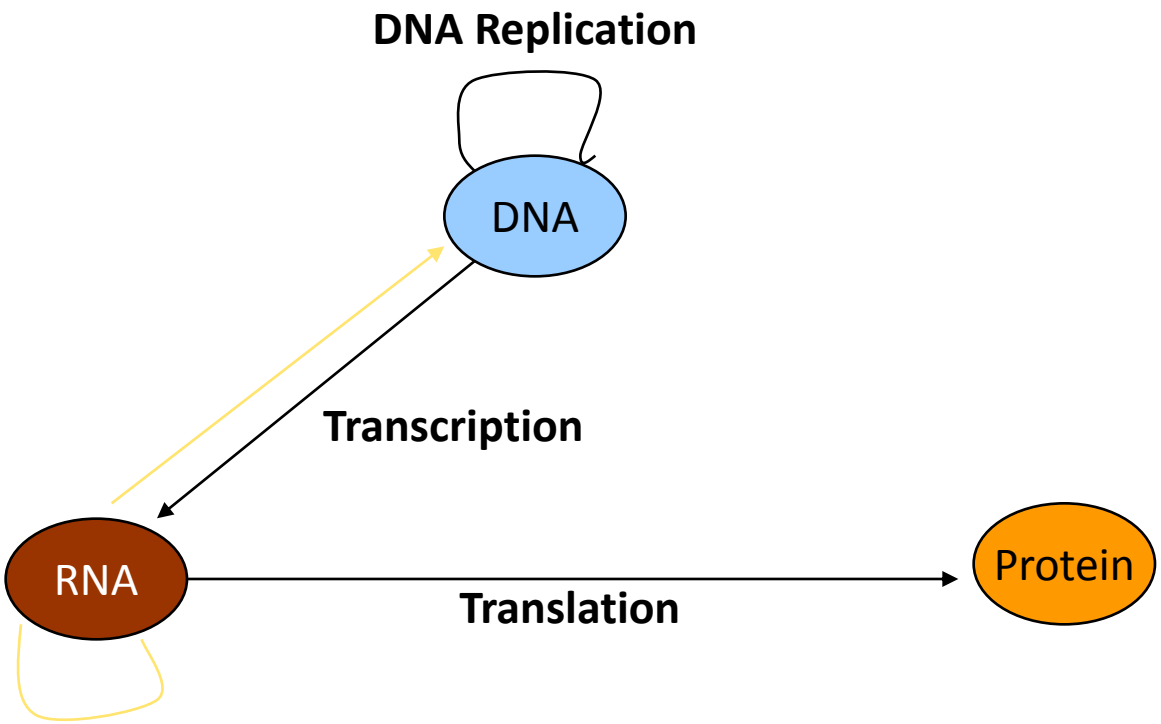
- DNA , a ‘text’ string on the four letter alphabet A, T, C and G ,spells out the biological information needed by the organism to synthesize all its proteins . Organisms differ from each other because the respective DNA molecules have different linear sequences of nucleotides which are responsible to 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

Proteins: Workhorses of the Cell

- 20 different **amino acids**
  - different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.
- Proteins do all essential work for the cell
  - build cellular structures
  - digest nutrients
  - execute metabolic functions
  - Mediate information flow within a cell and among cellular communities.
- Proteins work together with other proteins or nucleic acids as "molecular machines"
  - structures that fit together and function in highly specific, lock-and-key ways.

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.

(More about protein later)

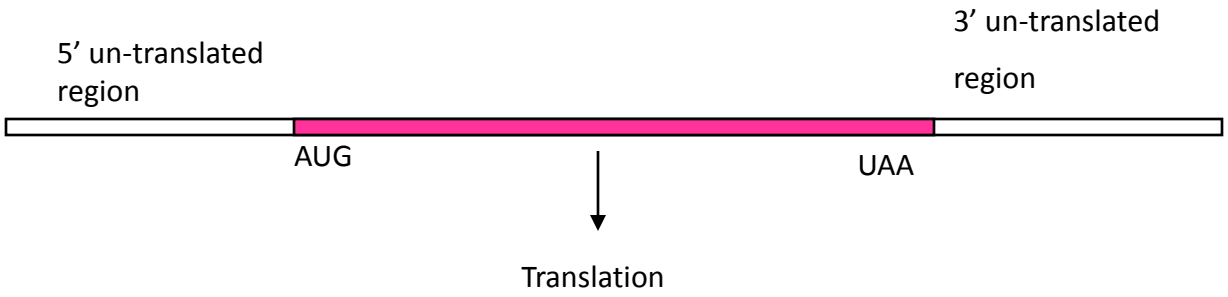
Gene expression

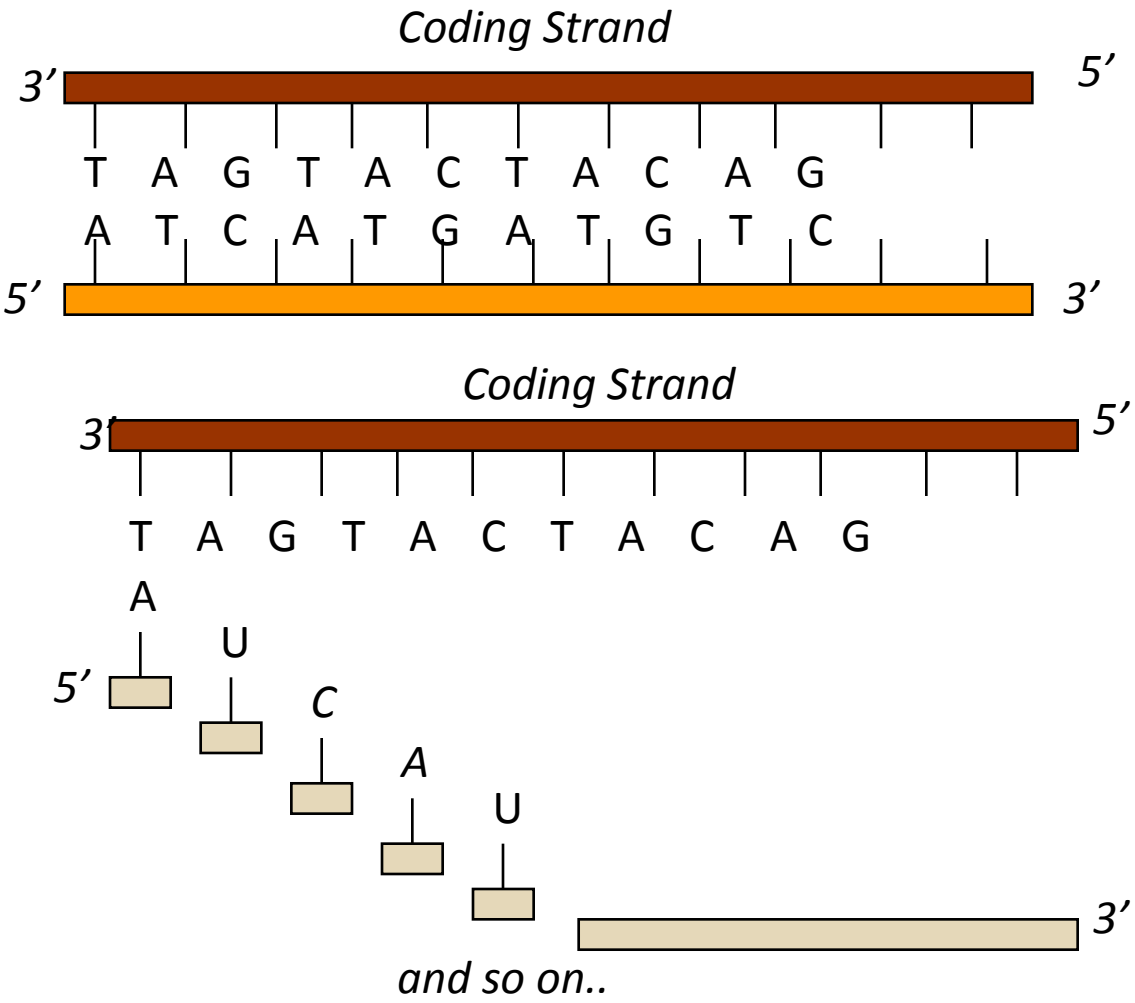
We will start with the simple situation of transcription of a single gene to a RNA. During transcription, the RNA transcript is built up in a step by step fashion using the DNA template as a guide. The template is read in 3'-5' direction and the RNA synthesis takes place almost like the DNA duplication 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



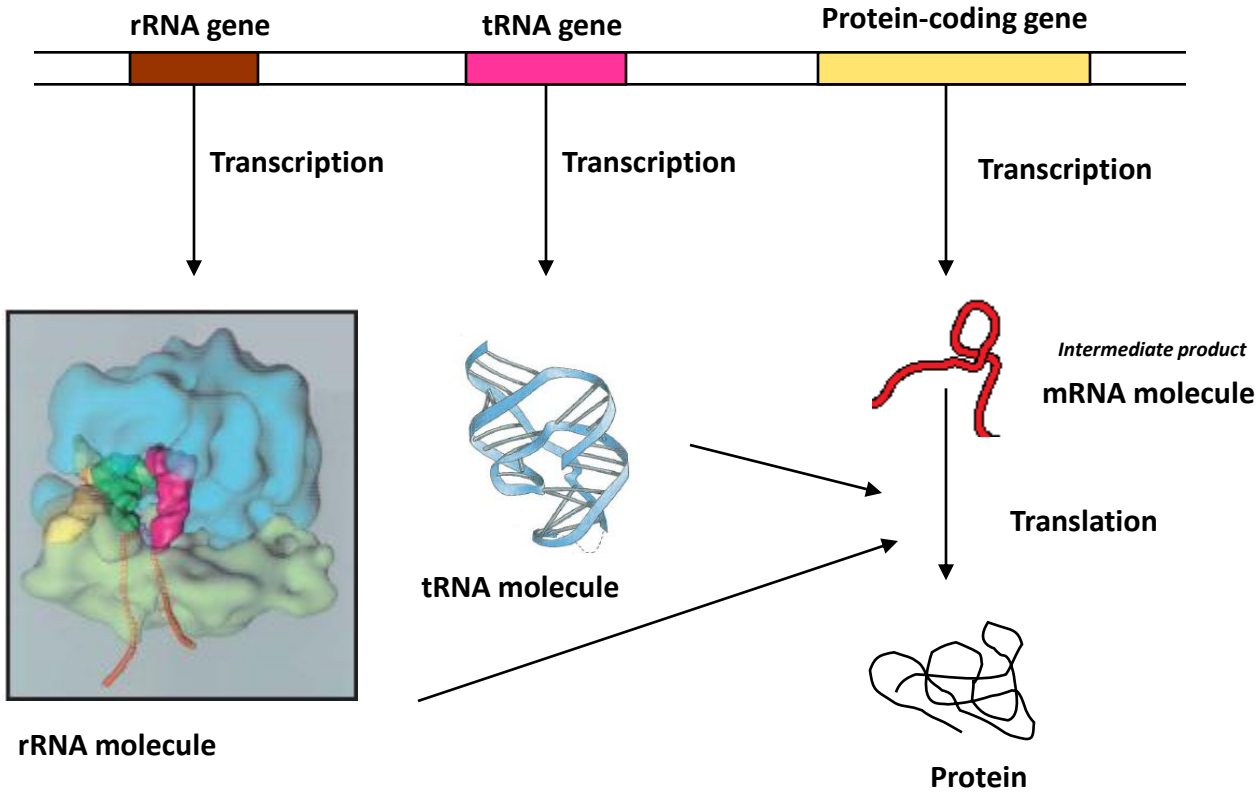


The final one-stranded RNA : AUCAUGAUGUC

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



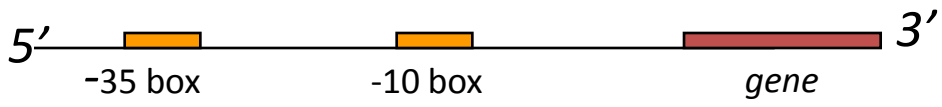
We will explain the functions of these RNA by directly quoting from the seminal text book "Cell". We will add some detail information derived from both of our reference text.

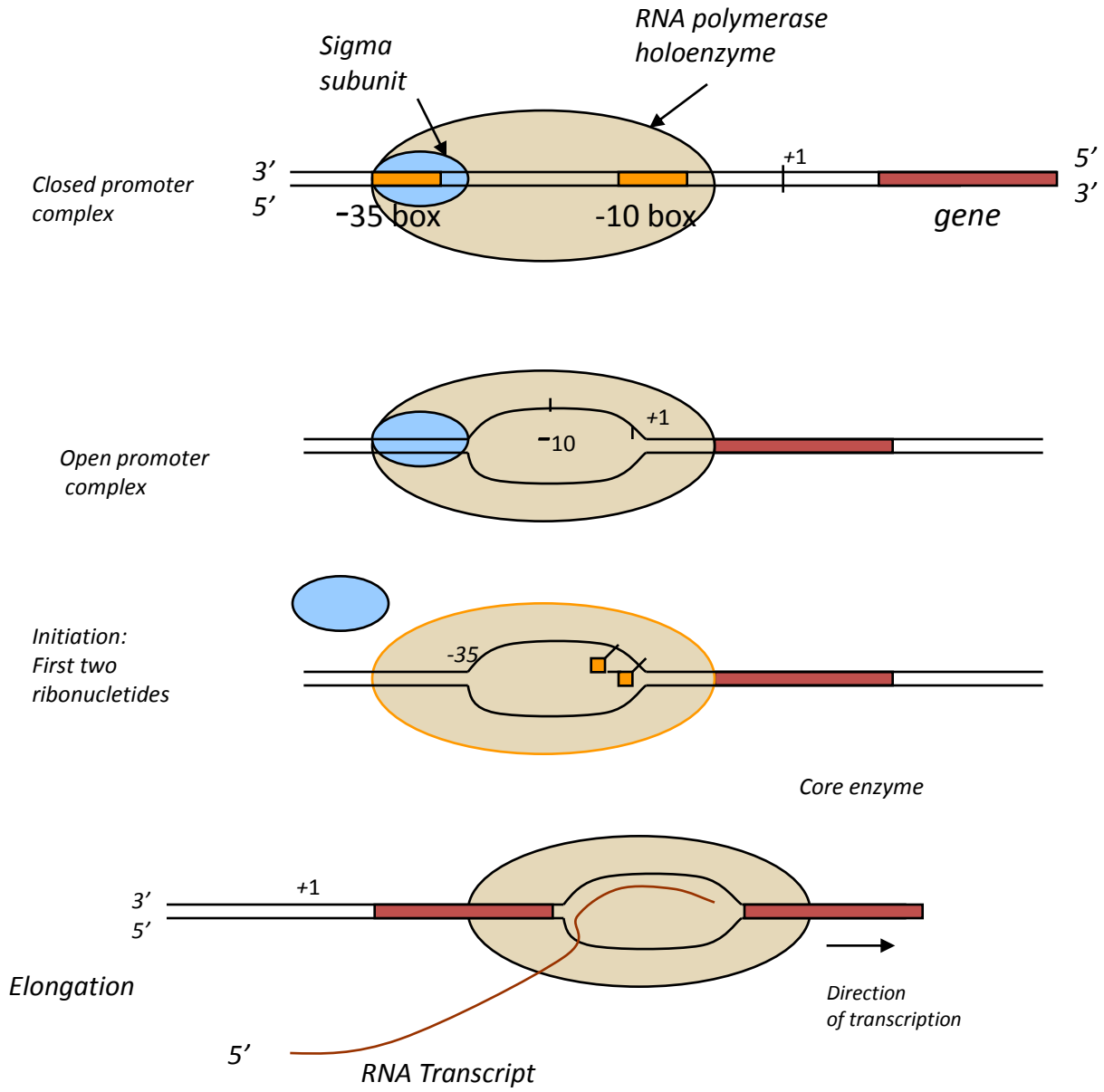
Transcription for E.Coli

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'




Sigma subunit
RNA polymerase holoenzyme
Closed promoter complex
3'
5'
-35 box
-10 box
+1
gene
5'
3'
Open promoter complex
-10
+1
Initiation: First two ribonucletides
-35
Core enzyme
3'
5'
+1
Elongation
Direction of transcription
5'
RNA Transcript

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)

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.

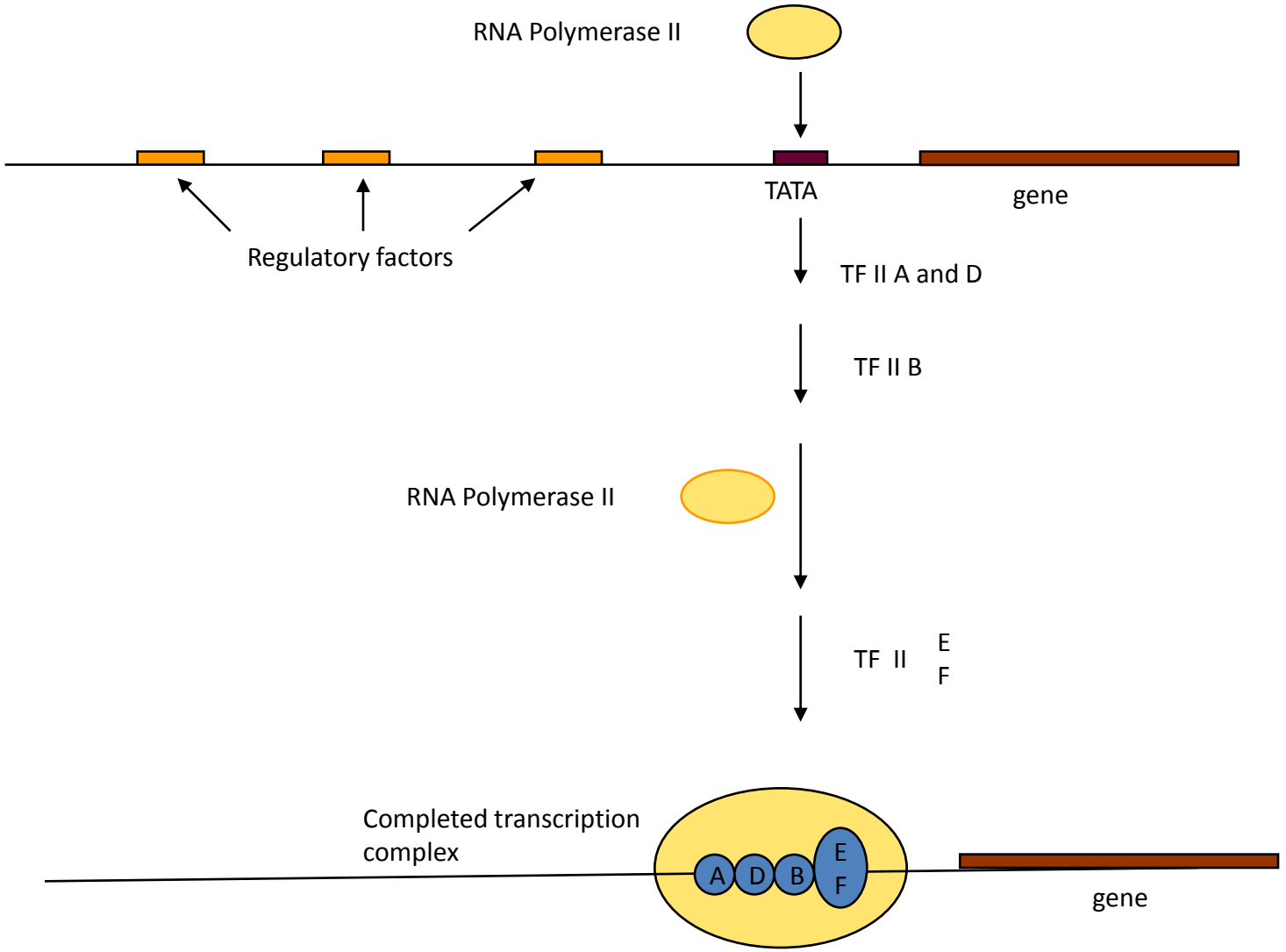
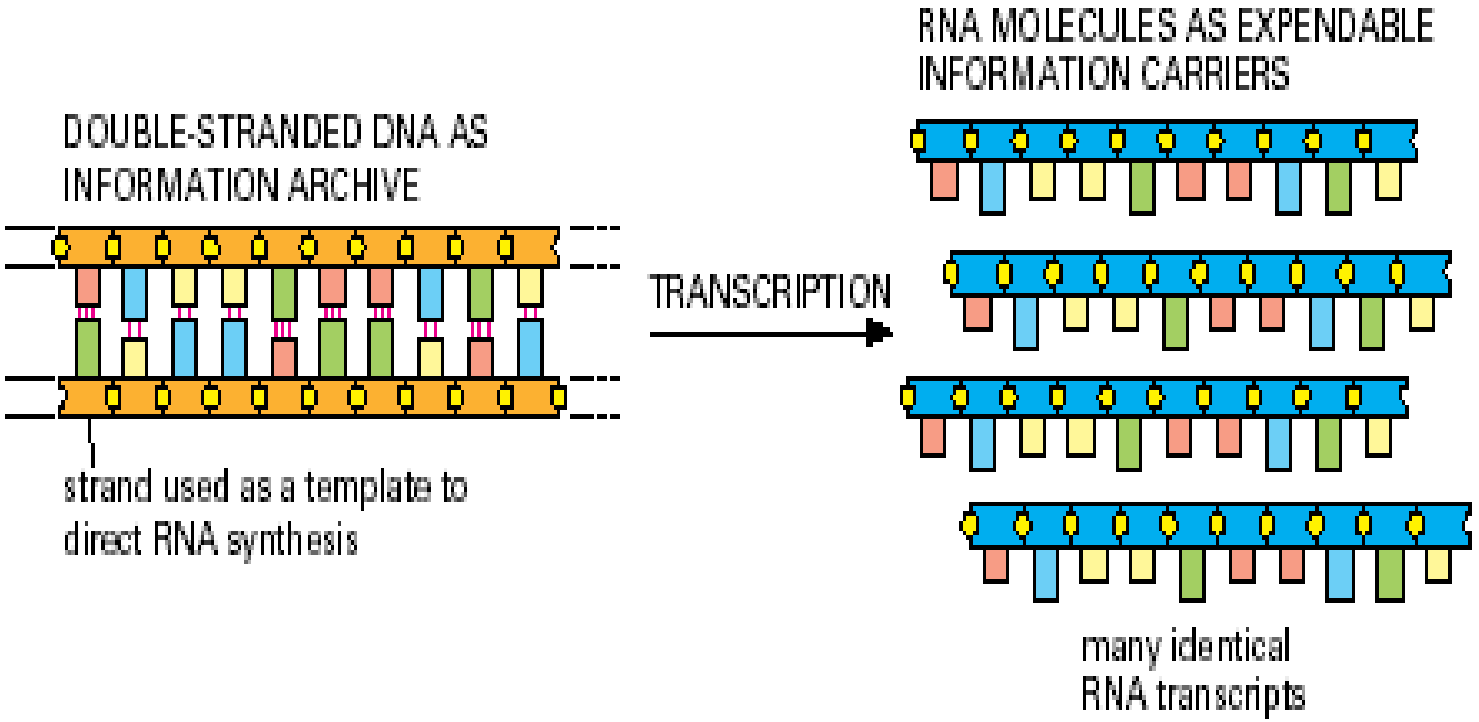

RNA Polymerase II
TATA
gene
Regulatory factors
TF II A and D
TF II B
RNA Polymerase II
TF II E F
Completed transcription complex
A
D
B
E
F
gene

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.



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			

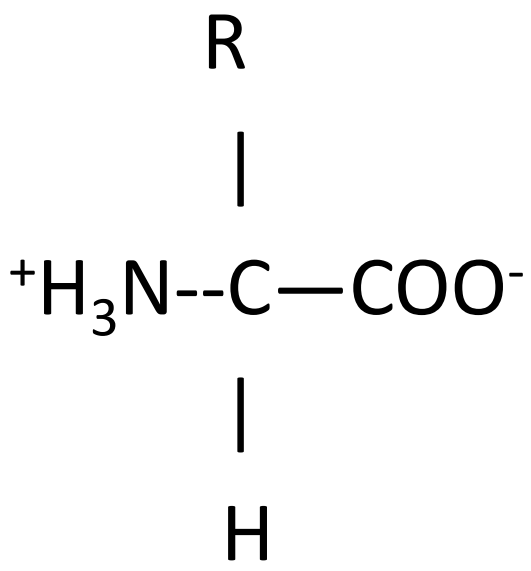
Proteins

- Complex organic molecules made up of amino acid subunits
- 20* different kinds of amino acids. Each has a 1 and 3 letter abbreviation.
- Proteins are often enzymes that catalyze reactions.
- Also called “poly-peptides”

*Some other amino acids exist but not in humans.

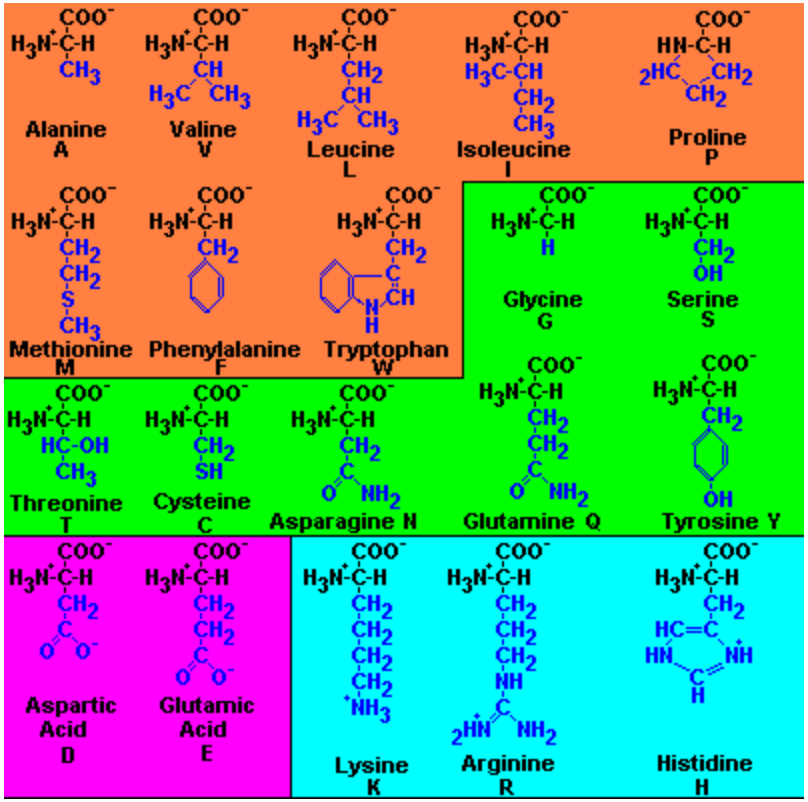
Proteins

- Composed of a chain of amino acids.

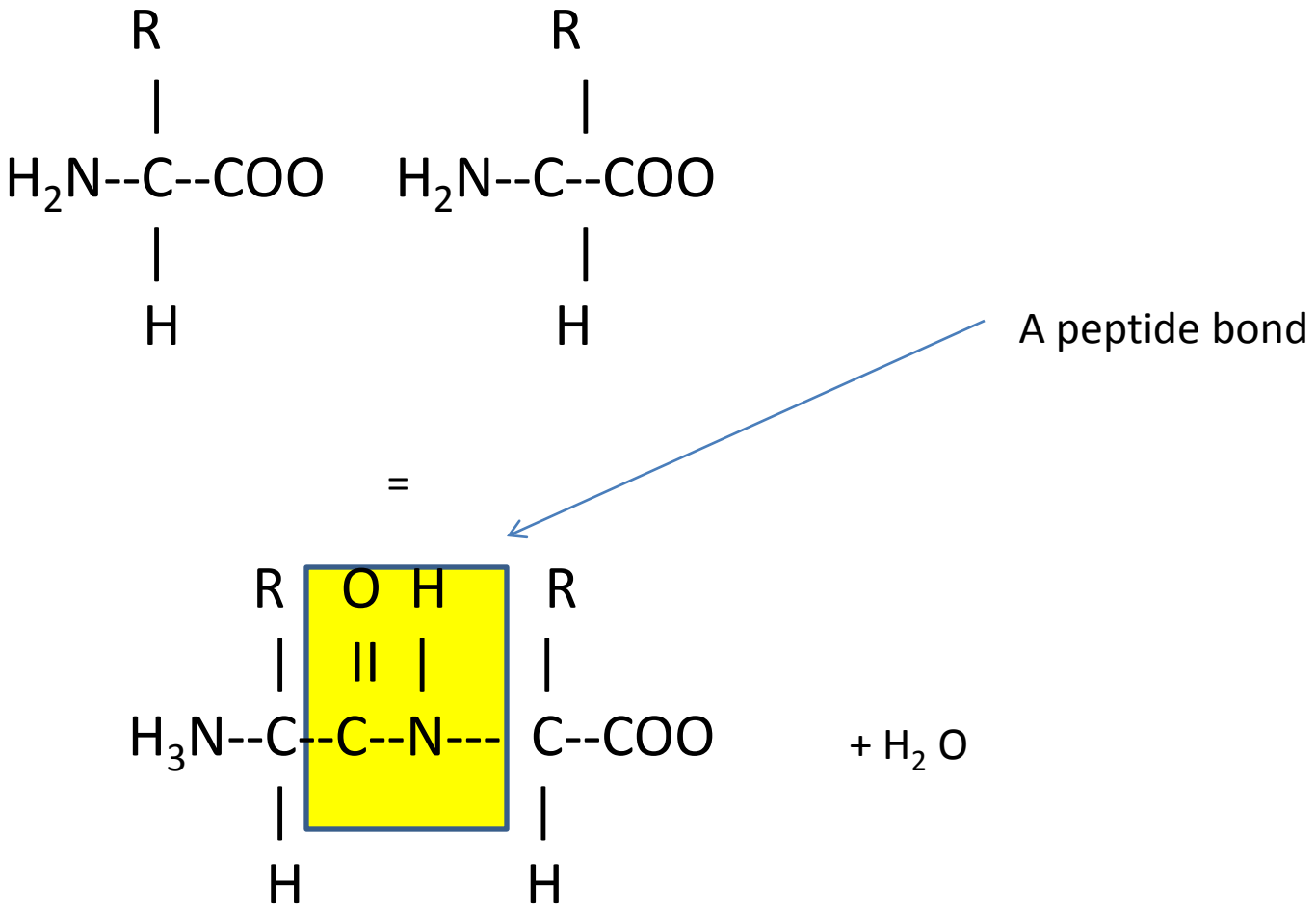


1. A hydrogen atom
2. A carboxyl)group ($-COO^-$)group
3. An Amino group (NH_3 +) group
4. The R group

20 amino acids



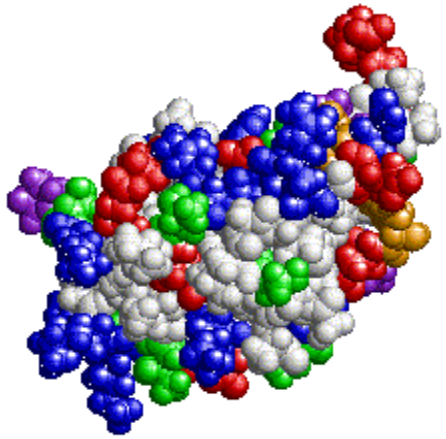
A peptide bond: Condensation of two amino acids



A chain of such bonds with different R form a protein. The last peptide is terminated by COOH

Protein structure

- Linear sequence of amino acids folds to form a complex 3-D structure.
- The structure of a protein is intimately connected to its function.



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 .