LECTURE NOTES FOR GENE STRUCTURE AND FUNCTION

May 17, 1990

Maxine F. Singer

Adapted from Dealing with Genes
by Paul Berg and Maxine Singer
Publication expected 1990/1991
University Science Books, Mill Valley, California

I. MOLECULES CONVEY INFORMATION

Three kinds of molecules are central to genetic processes: DNA, RNA and protein. The informational relationships between these genetic molecules are summarized in Figure 1. Genetic information in all cells is encoded in the variety and arrangement of nucleotides in their DNA. The information in a gene, a portion of a DNA molecule, is expressed first via the intermediary formation of a copy of DNA in the form of the related nucleic acid RNA (**transcription**), which in turn directs the production of specific proteins (**translation**). Each cell's and organism's characteristics are ultimately determined by the number and variety of proteins decoded from its DNA.

Fundamentally, information flows in one direction, from DNA to RNA to protein. Information can occasionally also flow from RNA back into DNA by a process called reverse transcription. DNA is transcribed into several kinds of RNA. One type of RNA (messenger RNA) is translated into proteins as described in Figure 1. Other kinds of RNA are involved in the chemical processes whereby cells manufacture proteins (Table 1).

Genetic continuity from one generation to the next requires that DNA fulfill another fundamental function besides encoding information. The information in DNA must be faithfully reproduced for delivery into new cells with each cycle of cell division. **DNA replication** is the process by which the parental DNA molecules are duplicated before being passed onto each of the offspring and must

occur with high fidelity.

A central feature of information transfer between nucleic acids, whether it be replication, transcription, or reverse transcription, is the involvement of the nucleic acid as a template to direct the assembly of new chains with related nucleotide sequences. The basic notion is that the order of nucleotides, A, T, G, and C, on an existing chain, the template, determines their order on the newly made chain. So far as is known, information stored in proteins is not used to assemble corresponding nucleic acids; thus, reverse translation is unknown. Nevertheless, proteins are critical participants in the processes that transfer information between nucleic acids and subsequently to proteins as well.

DNA. Double-helical DNA consists of a combination of two single DNA strands. The two chains are held together by weak chemical bonds called **hydrogen bonds** between atoms on the nucleotide bases on one strand and atoms on the nucleotide bases on the other (Figure 2). Only certain base pairs occur between chains: adenine (A) always pairs with thymine (T) and guanine (G) always pairs with cytosine (C). A consequence of these virtually invariant base pairs is that the sequence of bases on one strand uniquely defines the sequence of bases on the other strand to which it binds. The paired single DNA strands are

thus said to be **complementary** to one another. The complementary strands go in opposite directions. One chain goes from 5'-to-3' and the other 3'-to-5'. Note too that complementary strands generally have different nucleotide sequences.

The two complementary DNA strands in cellular DNA are wound around a common axis to form the double helix. Tracing the outside of the helix are long, repeating stretches of deoxyribose-phosphates. Pointing inward are the bases, with the weak hydrogen bonds between paired bases on the two strands holding the double helix together. Double helical DNA molecules are long, flexible, threadlike structures, with a nearly constant diameter, regardless of the order of the four nucleotides. DNAs differ in length, according to the number of base pairs. One thousand base pairs has a length of 0.00034 millimeters. Thus, for example, all the DNA in a human cell, about 6 billion base pairs, would be 2 meters long if it were stretched out. But in chromatin and chromosomes, the double helices are neatly wound and folded so that all the human DNA fits readily into a nucleus as small as 0.005 millimeters in diameter.

The DNAs in human chromosomes are long linear molecules when stretched out. Each chromosome appears to contain a single

DNA double helix. Other DNA molecules are circular and have no ends. Bacterial chromosomes, for example, are often circular and they are relatively small. The common bacterium, E. coli, has a single circular chromosome containing about 4 million base pairs and is thus about 1.4 millimeters in circumference. It is folded to fit in a cell no bigger than 0.001 millimeter across.

Usually, when discussing gene structure, DNA chains are represented horizontally, with only the sequence of nucleotides stated as in AAGCTG, etc. Direction is shown by indicating the 5' to 3' orientation: 5'-AAGCTG-3'. Thus, a double helical region and its complementary base sequence can be written as follows:

5'-AAGCTG-3'
3'-TTCGAC-5'

Because only relatively weak chemical bonds hold the two chains of the double helix together, input of rather small amounts of energy can unwind and separate the strands. Heating a solution of DNA in water close to the boiling point breaks the hydrogen bonds and unwinds the double helix without breaking the stronger bonds that hold the single strands together (Figure 3). Such DNA is said to be **denatured.** If the temperature is lowered the process is reversed. Under the right conditions, the two single strands realign properly and reform the original base pairs and the double helix - a process called **annealing** and often referred to as **hybridization.**

Unwinding and annealing of DNA strands is an artificial, laboratory reenactment of a process that is common and critical to the biological functions of DNA. It is also a critical component of the laboratory procedures used in studying and manipulating DNA. It is well to keep in mind that the fundamental requirement for annealing two DNA strands is that they be complementary. The order of As, Ts, Gs, and Cs on one strand must be matched by a corresponding order of the complementary Ts, As, Cs, and Gs on the other, with the two strands going in opposite directions.

RNA. RNA chains are very similar to single DNA strands. There are two major differences. First, the sugars in RNA are ribose, rather than deoxyribose (ribose contains one less oxygen atom than deoxyribose). Second, the base uracil (U) replaces thymine (T). RNA chains range in size from less than 100 to tens of thousands of nucleotides. The links that hold the nucleotides together are the same as in DNA: a phosphate links the 5' position of one ribose ring to the 3' position on the ribose of the neighboring nucleotide.

RNA forms the bulk of the nucleic acid in all cells, being five to ten times more abundant than DNA. As summarized in Table 1, there are several different kinds of RNA, most of which occur in all cells and play particular roles in translation, that is, the synthesis of proteins.

Unlike DNA, most cellular RNA is a single strand. During RNA synthesis, unlike DNA replication, only single RNA strands are made. However, within many RNA molecules there are short sequences of nucleotides which are complementary to other sequences in the same strand. Such pairs of complementary sequences can hydrogen bond together when they come in contact. For example, a 5'-UAUUC-3' sequence can pair with a 3'-AUAAG-5' sequence somewhere else in the strand, if the strand loops back on itself. Frequently, such intramolecular folding is critical to RNA function. The best understood structures are the transfer RNAs; a diagram showing the folded structure of a transfer RNA is shown in Figure 4.

If an RNA strand and a single DNA strand have complementary base sequences they can form an RNA-DNA hybrid double helix when conditions are appropriate. Here the U's in RNA pair with the A's in DNA while the T's in DNA pair with RNA A's. Such hybrid pairing is an essential tool in the laboratory manipulation of nucleic acids. For example, by measuring the ability of an isolated RNA to form a hybrid helix with a piece of DNA, the extent of their complementarity can be estimated. Also, a DNA segment can be used as a probe to detect the

presence of a complementary RNA in a mixture of RNAs (and vice versa).

Figure 5 illustrates how the complementarity between DNA and RNA chains can be estimated. First, the DNA double helix is unwound (denatured) with heat. RNA is added to the solution of DNA and the temperature is lowered. During annealing, base pairs form between complementary regions on the RNA and DNA. Unpaired, noncomplementary regions are removed by an enzyme that degrades single strands but not double helical regions. The length of the remaining double helix indicates the extent of complementarity.

Proteins. Proteins are the principal determinants of an organism's characteristics. They comprise the enzymatic machinery for the metabolic, energetic and biosynthetic activities of all cells as well as the regulatory elements that coordinate these activities. Proteins also contribute many of the structural elements that determine the shape and motility of cells and organisms. In short, organisms are what they are because of the array of proteins they manufacture.

At about the same time as the structure of the double helix was discovered, the final steps required to understand the basic structure of proteins were taken. Proteins contain one or

more **polypeptide** chains which are, like nucleic acids, long polymers. However, the individual molecules that are the building blocks for polypeptides are **amino acids**, not nucleotides. There are 20 different amino acids commonly found in protein in all living things (Figure 6). All these amino acids have one common structural group of atoms (NH_3CHCOO) and each amino acid has one unique chemical grouping called a side chain. The polypeptide chain is held together by strong chemical bonds between the carbon atom in the COOH (carboxyl) group on one amino acid and the nitrogen atom in the NH_2 (amino) group in the neighboring amino acid. Such bonds are termed **peptide** bonds, thus the term polypeptide (Figure 7). Like nucleic acids, polypeptide chains have a direction. One end contains a free amino group, the amino terminus, and the other a free carboxy group, the carboxy terminus.

Polypeptides comes in many sizes, generally between 50 and several thousand amino acids. But regardless of length, organism, or cell type, polypeptides and thus all proteins are constructed from the repertory of only 20 different amino acids. Each polypeptide possesses a different sequence of these amino acids along its length (Figure 8). This unique amino acid sequence, called the **primary structure**, directs the folding of the polypeptide into a characteristic shape. Strictly speaking, the term protein refers to a folded polypeptide.

Often, the proper three-dimensional form of a protein depends on the interaction of several polypeptide chains, which may be either identical or different in primary structure. Most important, it is this three-dimensional form that is responsible for the biological function of a protein and thus for its special role in a cell.

Sickle cell anemia is one of many convincing examples of the dependence of normal protein function on the correct primary and consequently the correct three-dimensional structure. A person with this genetic disease has hemoglobin that differs from normal hemoglobin by just one amino acid. Hemoglobin contains 4 polypeptide chains, 2 each of two different polypeptide chains called alpha and beta. The single amino acid change that causes sickling occurs in the beta chains, which are 121 amino acids long. This change from one amino acid to another profoundly alters the normal interactions between the chains. As a consequence, the shape of red blood cells is altered and blood flow through capillaries and small veins is impeded or interrupted. The replacement of the single amino acid reflects a change of a single DNA base pair in the gene for the beta chain, a mutation with a simple effect on a primary structure but a profound effect on the three-dimensional structure of hemoglobin. In the most general sense, it is this

relation between genes and protein structure that links an organism's properties to its genes.

II. TRANSLATING GENES INTO TRAITS

The story of sickle cell disease reveals the basic relation between gene structure and biological function. In the 1950's, the direct demonstration that mutations in a particular gene cause a change in the amino acid sequence of the protein encoded by the gene, supported the growing concept: genes specify amino acid sequence. The concept itself originated in the idea: one gene - one polypeptide. Confirmation of the concept and sophisticated insight into the relation between DNA sequence and protein structure, however, required an experimental approach that was virtually impossible with complex organisms like humans. Prokaryotes, mainly the bacteria E. coli, provided the necessary experimental system for combined studies on genetics and biochemistry.

A major advance came from analyzing multiple, independent mutations affecting an enzyme required by E. coli for synthesis of the amino acid tryptophan. This showed that positions of the mutations in DNA were in the same linear order as the amino acid changes in the protein. Thus, the linear order of the four bases in DNA specify the linear order of the twenty amino acids

in a polypeptide chain. The implied informational relationship between nucleotide and amino acid sequences was called the genetic code. The challenge was to decipher the code and to learn how it translates DNA into protein. To understand the structure of the code, which turned out to be virtually universal in living things, consider, first, the way genetic messages are expressed.

DNA Begets RNA. The expression of all prokaryotic and eukaryotic cellular genes begins with **transcription**, the process by which the nucleotide sequence in a gene's DNA is copied into a RNA chain. Transcription is carried out by **RNA polymerases**. These enzymes assemble RNA chains from individual nucleotides. The order of the ribonucleotides is specified by the sequence of deoxynucleotides in the DNA (Figure 9). Thus, a DNA strand serves as a template for RNA synthesis. During transcription, an RNA polymerase interacts with discrete sequences that define the beginning of each gene (**promoters**). The enzyme separates the two strands of the DNA duplex and, with the help of signals in the promoter, selects one DNA strand as the template for copying. Each nucleotide to be added to the RNA chain is determined by complementary base pairing to the successive nucleotides in the selected DNA strand. As the RNA polymerase moves from the beginning to the end of the gene's coding sequence, each properly matched nucleotide is added to the growing end of the

RNA chain. The newly made RNA strand has the same nucleotide sequence and direction as the non-template DNA strand. Specific types of nucleotide sequences signal transcription termination and release of the completed RNA. Mechanistically, the DNA-directed assembly of RNA chains is akin to DNA-directed DNA synthesis during DNA replication. The principal differences are 1) only one of the two DNA strands is copied into RNA, 2) the RNA nucleotides' sugars are ribose instead of deoxyribose and 3) the base uracil (U) replaces thymine (T) in the nucleotide sequence.

Transcription generates different kinds of RNA (Table 1). Most genes are transcribed into **messenger RNAs**, which encode the amino acid sequences of proteins and are used during translation. Within any cell, there are thousands of different messenger RNAs, each with a different sequence of As, Us, Gs, and Cs. Other genes are transcribed into other kinds of RNAs, the most abundant being the **ribosomal RNAs** and **transfer RNAs**. These are parts of the machinery that translates messenger RNA sequences into proteins but they are not themselves translated into proteins. Ribosomal and transfer RNAs occur in all cells. Although their structures vary some among different species, ribosomal RNAs have a fixed base sequence in each species; the thousands of copies in each cell are identical. Transfer RNAs, in contrast, are a mixture of

chains with different sequences, as described below. In prokaryotic organisms, a single **RNA polymerase** accounts for the production of all types of RNA. But eukaryotes use three distinctive kinds of RNA polymerase to make the three different types of RNAs: messenger, ribosomal, and transfer. Use of the three enzymes provides refined mechanisms for regulating the production of the different RNAs.

The Genetic Code. By the mid 1950's it was clear that the nucleic acid coding unit, the **codon**, was most likely three adjacent nucleotides in RNA, and that consecutive nucleotide triplets coded for adjacent amino acids in proteins. By 1964, the entire genetic coding dictionary was known. Each codon comprises three adjacent nucleotides in a DNA chain or its messenger RNA copy (Figure 10). Sixty-one of the possible 64 triplets each encodes only one amino acid. One of these triplets, ATG in DNA or the equivalent AUG in RNA, has a dual function. It encodes the amino acid methionine and it also marks the beginning of protein coding stretches - the start codon. The remaining three triplets TAG (UAG), TAA (UAA), and TGA (UGA) do not specify any amino acid, but any one of them signals the end of a protein coding sequence - a stop codon.

The code is said to be **"degenerate"** because more than one codon can specify the same amino acid; but the code is not ambiguous since a particular codon never specifies more than one

amino acid. With the knowledge of the codon dictionary, it is a straight forward exercise to translate on paper any DNA or RNA nucleotide sequence into its corresponding protein product. Usually, only one of the two DNA strands contains an informative array of codons in any particular region of a long double helix. The sequence of the complementary strand is "nonsense". Note, however, that it is the nonsense strand that is the template for messenger RNA synthesis. Synthesis produces the complement of the template, that is the informative sequence. Occasionally, the complementary strand may encode part of another gene.

RNA Begets Proteins. The elaborate process by which the sequence of nucleotides in a messenger RNA is translated into a polypeptide chain is complex and involves a very large number of repetitive steps. Physically, it occurs on intracellular particles called **ribosomes** that contain more than 50 different proteins and 3 or 4 different kinds of RNA molecules, ribosomal RNAs. The translation of messenger RNA into protein in bacteria is illustrated in Figure 11. On the top line, the first ribosome has already engaged the messenger RNA (mRNA) strand and carries a short polypeptide representing the first part of the encoded message. Polypeptides and messenger RNAs are not drawn to scale. A few seconds later (second line) the ribosome has moved along the messenger RNA which directs, triplet codon by triplet codon, the addition of more amino acids

to the nascent polypeptide chain. Meanwhile, a second, third, and fourth ribosome have successively engaged the messenger RNA chain and initiated the assembly of additional polypeptide chains. Somewhat later (third line), the first ribosome has completed assembly of and released its polypeptide. The ribosome itself is released from the messenger RNA, a process that involves breaking up the particle into two parts. Other ribosomes are nearing completion of their polypeptides. The nascent polypeptides begin to fold into their active structural forms even before they are complete.

Enzymes and other proteins associated with ribosomes foster the translation of messenger RNAs into polypeptides. The assembly of a polypeptide chain begins with the attachment of a ribosome to a messenger RNA. The polypeptide chain is elongated one amino acid at a time as the ribosome moves along the messenger RNA one codon at a time. As illustrated in Figure 12, the coding sequence starts with an AUG start codon and ends with one of the three stop codons. Note that the messenger RNA is shown 5' to 3', left to right. The start codon, AUG, is always near the 5'end of the messenger RNA and translation proceeds 5' to 3'. The coding region is flanked by untranslated regions on its 5' and 3' sides.

The key element in translation, the conversion of the genetic information encoded in the triplet codons in messenger

RNA into specific amino acids, depends, as does all of genetic chemistry, on complementary base pairing. Each amino acid is attached to a special cognate RNA, a **transfer RNA** (Figure 4), that contains a triplet that is complementary to the amino acid's coding triplet on messenger RNA (an **anti-codon**). Base pairing between the codon on the messenger RNA and the anti-codon on the transfer RNA puts the right amino acid in place and facilitates the joining of the amino acids to the growing ends of the protein chains (Figure 12A). The AUG methionine codon is recognized by a special, initiator transfer RNA. The polypeptide grows from the amino to the carboxyl end. Thus, methionine is always the first amino acid added at the amino end. Each codon is recognized by interaction with the complementary anticodon on the cognate transfer RNA. One pass of a ribosome along the length of the messenger RNA's protein coding sequence produces on molecule of the protein. Many identical polypeptides are thus produced from one messenger RNA strand.

Because the transcription of a gene by RNA polymerase generally begins some number of nucleotides before the start codon and ends some number beyond the gene's stop codon, messenger RNAs contain noncoding nucleotide sequences at both ends. The translational apparatus recognizes the triplet start and stop codons. After translation begins at the appropriate AUG start codon, it proceeds one by one through the series of

codons and ends when it encounters a stop codon. This means that each coding region contains a multiple of three nucleotides because all codons, including start and stop have three nucleotides. Note that there are three possible frames on a messenger RNA, depending on which of a series of 3 nucleotides is chosen as the first in the first codon. The translational apparatus selects the proper frame - the reading frame - by recognizing the start codon - AUG. In the example in Figure 13, as in most cases, the two alternate frames (labeled B and C) are interrupted by stop codons and cannot be translated. Only frame A is "open" throughout.

Another important fact about translation is that reading frames make sense in only one direction along a DNA or messenger RNA chain. The "beginning" of a messenger RNA chain is that end whose deoxyribose has no neighbor on its 5' side. At the end of the messenger RNA, the last deoxyribose has no neighbor on it's 3' side. Such a direction is, by convention, said to be 5'- to - 3'. The AUG start codon is near the 5' end and specifies the amino terminus of the polypeptide, always the amino acid methionine. The stop codon is closer to the 3' end of the messenger RNA. The carboxy terminal amino acid of the polypeptide is specified by the codon immediately preceding the stop codon. This is somewhat confusing so perhaps a summary rule is helpful. Translation goes from amino to carboxyl on the polypeptide and from 5'-to-3' on the messenger RNA.

III. GENE STRUCTURE

Our first impressions of gene structure came from studies on bacteria. A bacterial gene, a stretch of DNA ranging from tens to thousands of base pairs in length, has a structure whose sequence has a simple relation to the messenger RNA: a double-helical stretch of DNA in which a coding region is flanked by sequences that are transcribed but not translated. One strand of the DNA, the template, has the same base sequence as the messenger RNA.

Long before human gene structure could be studied directly, it was known that the genetic systems of bacteria and complex organisms share certain fundamental properties. Their genetic material is DNA, the DNA is replicated in similar ways, genetic information flows from DNA to RNA to protein, and the genetic code is the same. It was expected, correctly as we now know, that complex organisms, especially multicellular ones, would have more genes than prokaryotes. And it was widely assumed, incorrectly, that the basic structure of the genes in the two types of organisms would be similar. These assumptions were demolished almost as soon as the first mammalians genes were analyzed.

The protein coding sequences in a mammalian gene are not necessarily in a single contiguous stretch of DNA as they are in a bacterial gene. Instead, coding regions are often discontinuous, being interrupted by stretches of non-coding DNA; such non-coding, interrupting DNA segments are called **introns**. The coding segments of genes, are referred to as **exons** (Figure 14).

The first inklings of the existence of introns emerged from studies of animal viruses. Surprisingly, the nucleotide sequences of various virus messenger RNAs did not line up with the corresponding nucleotide sequences in the virus' DNA. Indeed, the messenger RNAs seemed to be composed of base sequences present in discontinuous stretches in the virus DNA. Similar findings were made in mammalian cellular genes.

Now we know that most (though not all) of the genes that code for polypeptides in vertebrates (including humans) and plants have at least one intron and some have very many more (Table 2). In some organisms, the amount of DNA in the introns can exceed the amount within the exons by as much as ten-fold.

It is helpful to remember an important difference between bacteria and complex organisms (**eukaryotes**). In bacteria, there is no barrier between the site of transcription on DNA and the ribosomes necessary for translation. Ribosomes bind to

messenger RNA and begin protein synthesis even before transcription is completed. However in eukaryotes, messenger RNA and transfer RNA and ribosomal RNA are transcribed from a DNA template that is in the nucleus. The RNAs must be transported across the nuclear membrane into the surrounding cytoplasm, the site of protein synthesis. Early work on the RNA within the nuclei of eukaryotic cells produced some puzzling observations. The RNA molecules were, on average, much longer than expected and were very heterogeneous in size. Moreover, much of the RNA did not appear to be transported out of the nucleus.

The existence of introns and the enormous variation in their number and size explained these oddities. The heterogeneous nuclear RNA is a mixture of transcripts of many chromosomal genes. When a gene is transcribed, RNA polymerase begins to copy the DNA at the position corresponding to the start (the 5' end) of the messenger RNA. It continues transcribing through exons and introns until it reaches the end of the gene (Figure 14). Therefore, the heterogeneous nuclear RNAs are RNA strands containing both exons and introns. In contrast, the mature messenger RNAs in the cytoplasm are completely free of introns. In fact, only intronless messenger RNAs leave the nucleus for the cytoplasm. The complex mixture of RNA transcripts inside the nucleus is not only heterogeneous

but most of it never leaves the nucleus because it is intron RNA.

Introns are removed from the RNA transcripts made in the nucleus by a mechanism called **splicing.** In the process, each intron is cut at its ends and the two flanking exons are joined together to generate a continuous stretch of exons that contain the coding sequence (Figure 14). Although simple in design, an elaborate cellular machinery is needed to ensure the fidelity of splicing. For only if the cleavages at the two sides of an intron are precise and the exons are joined to each other intact and in the correct order, can the messenger RNA be translated into its intended protein product. Not a single nucleotide in the exon can be lost.

Small nuclear particles composed of special RNAs (see Table 1) and proteins catalyze the splicing process in most instances. They recognize characteristic short nucleotide sequences that occur at the two ends and within all introns. Remarkably, some introns in some organisms can splice themselves out of their RNAs without the help of the special particles. The intron folds itself up in a special way, so that cutting at the intron boundaries and joining of the exons occurs spontaneously, without outside intervention. Such self splicing is generally believed to be the evolutionary forerunner of

present day particle-mediated splicing.

Most often, splicing results in the excision of all the introns and the conservation of all the exons in their original order as a continuous sequence in a single messenger RNA (Figure 15). However, sometimes splicing occurs in more than one way. An exon can become an intron. Consequently, several different messenger RNAs are formed from the same initial transcript (Figure 15). Each has a different subset of exons and, therefore, each encodes a structurally different protein. Such alternative splicing is of considerable importance. It provides an efficient way of diversifying the proteins produced from a single gene, without altering the gene's structure in the DNA. Moreover, alternative splicing can be regulated so that a gene can be expressed one way at one time during development or in one kind of tissue, and another way at a different developmental stage or in a different tissue.

RNA Maturation. Most of the transcripts produced in eukaryotic cell nuclei require alteration before they become functional RNA molecules. After being transcribed, spliced, and modified, if necessary, messenger RNA, transfer RNA, and ribosomal RNA participate in protein synthesis in the

cytoplasm. Thus, in addition to being altered, they must be transported across the nuclear membrane into the cytoplasm before they can carry out their proper cellular roles. A few small RNAs remain in the nucleus where they are required for a variety of processes such as ribosomal RNA processing, intron splicing, and transcription termination.

The maturation of a transcript of a protein coding gene into a messenger RNA requires several steps (Figure 16). As already emphasized, many such RNAs must be spliced to remove introns. In addition, eukaryotic messenger RNAs always have a special chemical modification, referred to as a "cap", at their 5' end (Figure 16). Caps facilitate the binding of ribosomes to messenger RNA, thereby increasing the efficiency of translation. The third modification associated with messenger RNA maturation occurs at the 3' end of the RNA molecule. Transcription generally proceeds well beyond the 3' end of the gene. The transcript is then cleaved about 20 nucleotides past a specific sequence 5'-AAUAAA-3' that always occurs after the end of the coding region. At the cleaveage site, which is then the 3' end of the RNA, anywhere from 50 to 200 adenine nucleotides (As) are added to form what is called a polyA tail. Thus, the functional messenger RNA carries a noncoding tail that was not present in the gene.

IV. SWITCHING GENES ON AND OFF

Learning how genes are transcribed and translated is only one aspect of the study of gene expression. Another aspect is the control regulating which gene(s) is to be expressed and the rate and amount of gene expression under a wide variety of circumstances.

Gene expression is most frequently regulated at the level of transcription, that is, of messenger RNA production. Generally, initiation of transcription is the regulated event. Whether transcription of a particular gene is on or off is determined by proteins that bind, reversibly, to special DNA sequences in the neighborhood of the gene. Interaction of these regulatory proteins with the DNA can have either a negative or a positive effect on transcription. Accordingly, the proteins are called transcriptional **repressor** or **activating** factors, respectively. To appreciate such regulatory mechanisms it is best, before considering complex organisms, to look first at the situation in the bacteria Escherichia coli, because it is more clearly understood. We begin with a well studied example.

Regulation of bacterial gene expression. The E. coli enzyme beta-galactosidase breaks down the sugar lactose into

two simpler sugars, glucose and galactose. If E. coli is grown in the presence of glucose as a nutrient, it does not synthesize beta-galactosidase. But the enzyme is produced if lactose is the only sugar available. Several DNA sequences that precede the 5' end of the beta-galactosidase coding region on E. coli DNA are involved in the regulation (Figure 19). One, the promoter, binds the RNA polymerase enzyme that will transcribe the gene. A second sequence, the operator, overlaps the promoter sequence and is closer to the beta-galactosidase gene. The operator sequence interacts with another protein, the repressor. Because the binding sites for the repressor and RNA polymerase overlap on the DNA, binding of repressor blocks RNA polymerase from binding to the promoter, and no transcription occurs. However, if lactose is supplied to the E. coli, the sugar binds specifically to the repressor protein thereby altering the three-dimensional structure of the repressor so that it no longer binds to the DNA. RNA polymerase can now find the promoter and begin transcription. This, like the case of sickle cell hemoglobin, illustrates the importance of protein folding to biological activity. Note too that the beta-galactosidase gene repressor specifically recognizes and binds to the short sequence of DNA nucleotides in the operator. It does not bind to other DNA sequences. Consequently, the repressor only regulates the beta-galactosidase gene. Other genes in E. coli are not affected by the beta-galactosidase repressor.

Besides the negative control provided by repressor, beta-galactosidase gene expression is also subject to positive regulation. Transcription can only start if a distinct activator is present. The activator is a protein called CAP, bound to a small molecule called (cAMP). The complex of CAP-cAMP binds to still another short DNA segment on the 5' side (leftwards) of the promoter-operator region. CAP-cAMP binding is required for RNA polymerase action. The CAP protein itself is only in the correct three-dimensional form to bind to the DNA if there is no glucose available to the bacterium. Thus, expression of the beta-galactosidase gene depends on two environmental conditions - the presence of lactose and the absence of glucose.

Transcription of most bacterial genes is subject to analogous, complex regulation. But repressor and activator proteins are not the sole means for regulating gene transcription. In some instances, feed-back systems operate so that the protein product of a gene's expression is itself the regulator of the gene's transcription RNA production can also be regulated by controlling the growth rates of RNA chains rather than the rate of initiation, or by determining whether transcription continues through the entire gene or is terminated at specific stop-go signals within the gene.

Gene expression can also be regulated during the translation of messenger RNA into polypeptide. Here too, control is most often exercised at the initiation step, the reading of the first codon, although, later steps in polypeptide chain assembly can also be regulated. Additional regulatory events can occur during conversion of a completed polypeptide into a functional protein because many polypeptides are modified by the addition of special chemical groups. Each such modification is catalyzed by one or more enzymes whose abundance or level of activity can itself be modulated. Moreover, many proteins must be transported to a particular cellular location in order to function properly. For example, the proteins in a ribosome must be transported to the site of ribosome construction. Others must be delivered to the cell membrane or even to the outside of the cell. Control of the traffic, the transport of proteins to their proper sites, also regulates the levels of functional proteins. Thus, numerous diverse and complex mechanisms cooperate to regulate the conversion of a gene to a functional gene product. Often, the processes are unique for a particular gene, for the physiological condition of the cell, or for the external environment.

Sophisticated switches in eukaryotes. Virtually every cell in a complex organism contains the same DNA - the cellular genome. Some genes actively produce their gene products in all cells. For example, ribosomes occur in all cells and ribosomal

RNA and protein genes work in all cells. Such genes are often called "housekeeping" genes, because they are involved in the basic functions common to all living cells. However, many genes are active only in specific cells or tissues. The visual pigment genes are a case in point. They are expressed only in the cells of the retina and even there, rod cells express the rhodopsin gene and cone cells each express only one of the several pigments for color perception (i.e., green, red, or blue sensitive). Yet, the thousands of different cell types and the many specialized tissues that they form, all arise from a single cell, the fertilized egg cell. Differential gene expression imparts to these cell types their unique shapes and functions. Through the successive cell divisions, starting with a fertilized egg, cells arise in which certain genes are turned "on" and other turned "off". Moreover, the on/off switches are activated at very precise times during development. In addition, the position of a cell in the developing embryo influences which genes are on or off. Understanding these extraordinary events is one of the most challenging and interesting problems in biology.

How are the positional and temporal regulation of cell and tissue specific gene expression achieved? Primarily by controlling the initiation of transcription.

The analysis of eukaryotic gene structure and function revealed that the signals that regulate gene transcription are considerably more complex than bacterial regulatory signals. Three different kinds of regulatory elements are known in bacteria. One kind, a promoter includes the sequences that determine where transcription begins. In E. coli for instance, the promoters are defined by two short DNA nucleotide sequences that occur about 10 and 35 base pairs before the site of transcription initiation. A second kind of element marks the end of a gene or group of genes and triggers transcription termination. And, finally, there are DNA sequences around the promoter (such as operators) that are recognized by specific regulatory proteins such as repressors and activators to modulate transcription. All these regulatory sequences depend upon interactions with proteins to influence the expression of the neighboring coding sequences. Bacteria contain only a single kind of RNA polymerase which transcribes all types of genes.

In contrast, the DNA sequences that regulate eukaryotic gene expression occur in a variety of locations and even at various distances and directions relative to the transcription start and stop sites. Moreover, three different RNA polymerases, I, II and III, transcribe three distinct classes of genes and each class is associated with different and specific transcriptional control and terminator signals.

Genes for ribosomal RNA are transcribed by RNA polymerase I. The enzyme acts in conjunction with accessory regulatory factors (proteins) that recognize DNA nucleotide sequences on the 5' side of genes encoding ribosomal RNA. Some of these factors function in a species-specific manner. Ribosomal RNA genes from fruit flies, for example, are not transcribed by the RNA polymerase I system of human cells. In contrast, RNA polymerase II and III, regardless of their species of origin, transcribe appropriate genes from any species. These enzymes also depend on a variety of regulatory proteins. RNA polymerase II transcribes all genes encoding proteins and genes encoding some RNAs (e.g., small nuclear RNAs). RNA polymerase III transcribes genes encoding certain other RNAs (e.g., a small ribosomal RNA and transfer RNA).

Much of RNA polymerase specificity relies on multiple specifically organized DNA nucleotide sequences in the vicinity of the transcribed genes and their interaction with specific proteins. In this sense, the basic logic of transcription regulation is similar in bacteria and eukaryotes: it involves the precise recognition of specific DNA sequences by "matching", binding proteins. The proteins may either activate or inhibit gene transcription. While the transcriptional on/off switches are probably not absolute, experimental evidence suggests that rates of expression may be regulated over a million fold range.

Molecular genetics and biochemical studies are revealing the detailed mechanisms governing the differential regulation of gene expression. The regulatory signals in DNA consist of complex areas of relatively short DNA sequence motifs. Each motif is a binding site for a specific protein, a transcription factor. For example, genes that are uniquely transcribed in white blood cells, such as antibody genes, contain an array of motifs that are recognized by cognate transcription factors, some of which are restricted to white cells. Similarly, genes expressed only at certain times, or under particular environmental conditions, contain regulatory motifs that interact with cognate proteins that are only present or active at those times, or under those conditions. Turning a gene on or off depends on the particular assortment and arrangement of DNA sequence motifs, the availability of the cognate transcription factors, and the way the factors influence transcription initiation. Binding of multiple transcription factors in the gene's regulatory regions facilitates either the assembly of the relevant RNA polymerase into an active transcription complex, activation of such a complex, or both.

An important corollary of this general mechanism, is that different cell types have different active regulatory proteins, and there is ample evidence to support the model. But this explanation doesn't go very far because we now must explain what controls the presence or absence of the specific regulatory

proteins. Moreover, there are hints that more than specific regulatory proteins are involved. It is possible that relatively large chromosomal regions can be "opened" or "closed" for gene expression by changes in chromatin structure.

An extraordinary feature of the regulation of gene transcription by a combination of DNA sequence motifs and cognate protein transcription factors is its universality. Remarkably, corresponding transcription factors from such diverse sources as yeast, flies, and mammals are interchangeable. A yeast or fly transcription factor can interact with the corresponding mammalian DNA sequence motif, and with the other mammalian proteins required to form an active transcription complex. Transcription of the mammalian gene is turned on. Often, the reverse is also true. This universality implies a notable conservation of structure in the course of evolution.

The transcription of some genes is modulated by changes in DNA structure that do not alter the basic nucleotide sequence. One such change is associated with a chemical modification of the C bases in sequences 5'-CG-3' near the start (5'# end) of the gene. The modification introduces a methyl group (CH_3) on the ring of the cytosine base. The reaction is catalyzed by special methylase enzymes. Increased methylation near the beginning of a gene often correlates with reduced or no

gene transcription while reduced methylation correlates with high levels of expression. Another effect on transcription is associated with structural changes in chromatin itself. Although poorly characterized chemically, such changes can influence the expression of nearby genes.

The unique features of eukaryotic cells and the structure of their genes and genomes confer special opportunities for controlling the flow of genetic information. For example, a complete gene transcript is functionless unless its introns are properly spliced. Similarly, transcripts destined to become messenger RNAs must be modified at both ends with caps and polyA tails to be functional. Messenger RNAs must cross the nuclear membrane into the cytoplasm before they can be translated. Each of these events provides a juncture at which regulation may occur. Among the more unusual mechanisms that regulate gene expression are those which, like the formation of functional immune protein genes in vertebrates, involve the rearrangement of genomic DNA. In addition, polypeptide levels are modulated by controls on messenger RNA processing and stability and on protein modifications, transport, and stability of polypeptides.

V. WHAT IS A GENE?

A gene is no longer an abstract unit of heredity governing the nature of a particular trait as it was from Mendel's time until the middle of th is century. However, arriving at a single, consistent, new molecular definition is not simple. For example, is an intron part of the gene? In some instances sequences that are an intron for one gene contain another, independent gene. Are regulatory sequences part of a gene? Some of them are thousands of base pairs away from the gene they control. What about noncoding sequences that flank a gene's coding sequence and are transcribed, becoming part of the messenger RNA? Are they part of the gene? In fact, several different possible definitions are plausible, but no single one is entirely satisfactory or appropriate for every gene. At best, it is possible to describe the elements that usually occur in a functional unit of heredity. Before doing that we summarize a few terms.

First, most genes include DNA nucleotide sequences that encode proteins. These sequences will be translated from the corresponding messenger RNA by ribosomes and transfer RNA, according to the genetic code. Some genes do not encode proteins. They are transcribed into RNA molecules which, rather than serving as messenger RNAs, are themselves functional. Thus, there are genes that encode ribosomal RNAs and transfer RNAs.

DNA segments that encode all or a portion of a polypeptide or a functional RNA are called coding regions. Other DNA segments, that are not represented in the gene's product, are also part of the functional unit and are called non-coding segments. Non-coding segments include introns and various regulatory signals that flank genes. The terms 5' flanking sequence and 3' flanking sequence refer to nucleotide sequences that precede and follow (left and right) the coding regions, respectively. Although DNA is double stranded, usually only one strand contains coding sequences in the segments encompassing a gene and only one strand is transcribed.

A gene, a unit of heredity, is a combination of DNA segments that together constitute an expressible unit, expression leading to the formation of one or more functional molecules that may be either RNAs or polypeptides. The segments of a gene include first, the transcribed region or the transcription unit. The transcription unit encompasses the coding sequence (exon), any introns, any transcribed 5' and 3' flanking sequences that surround the ends of the coding sequences. These flanking sequences may include certain regulatory sequences. Second, a gene also includes the regulatory sequences that are required for specific gene expression but are not transcribed and not included in the RNA. It is important to recognize that a mutation that destroys a transcriptional regulatory sequence that is required for a

messenger RNA synthesis can wipe out a functional gene as readily as a mutation that destroys the reading frame. This notion emphasizes that the word mutation also now has a chemical definition: a change in DNA sequence that alters the expression or coding information of a gene.

Table 1 Some Important Cellular RNAs

Types of RNA	Approx. Number of Different Kinds in Cells	Approx. Length in Nucleotides	Distribution*
Transfer RNA	40-60	75-90	P,E
5S Ribosomal RNA	1-2	120	P,E
5.8S Ribosomal RNA	1	155	E
Small Ribosomal RNA	1	1600-1900	P,E
Large Ribosomal RNA	1	3200-5000	P,E
Messenger RNA	thousands	vary	P,E
Heterogeneous nuclear RNA	thousands	vary	E
Small cytoplasmic RNA	tens	90-330	P,E
Small nuclear RNA	tens	58-220	E

*P = prokaryotic, E = eukaryotic

Table 2

Introns in a Few Representative Human Genes*

Gene	Exons	Introns	
	total bp	number	total bp
erythropoietin	582	4	1,562
adenosine deaminase	1,500	11	30,000
low density lipoprotein receptor	5,100	17	40,000
Thyroglobulin	8,500	>40	100,000
clotting factor VIII	9,000	25	177,000

*The genes are named according to the protein they encode. The point here is to illustrate the diversity of gene structure and the large amount of DNA consumed by introns.

Figure 1

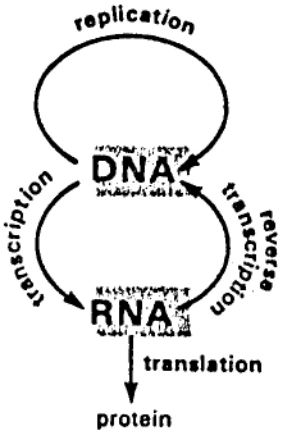


Figure 2

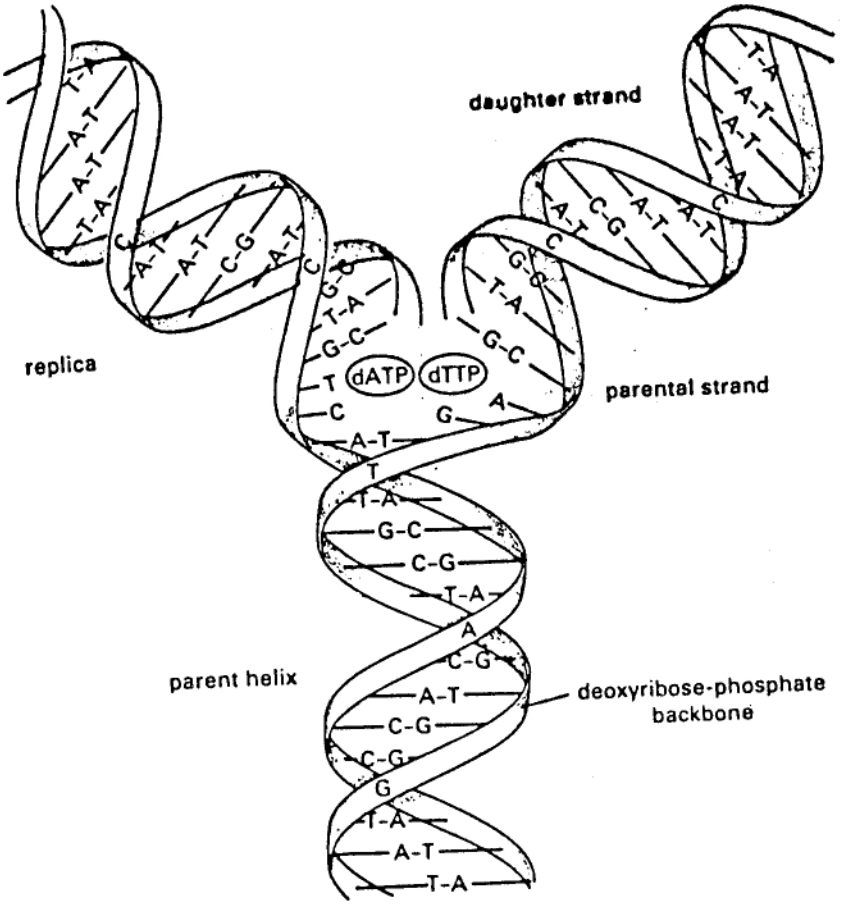


Figure [redacted] 3

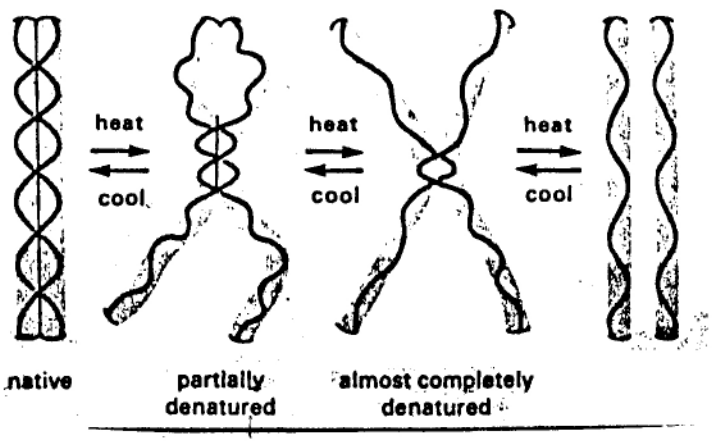


Figure 4

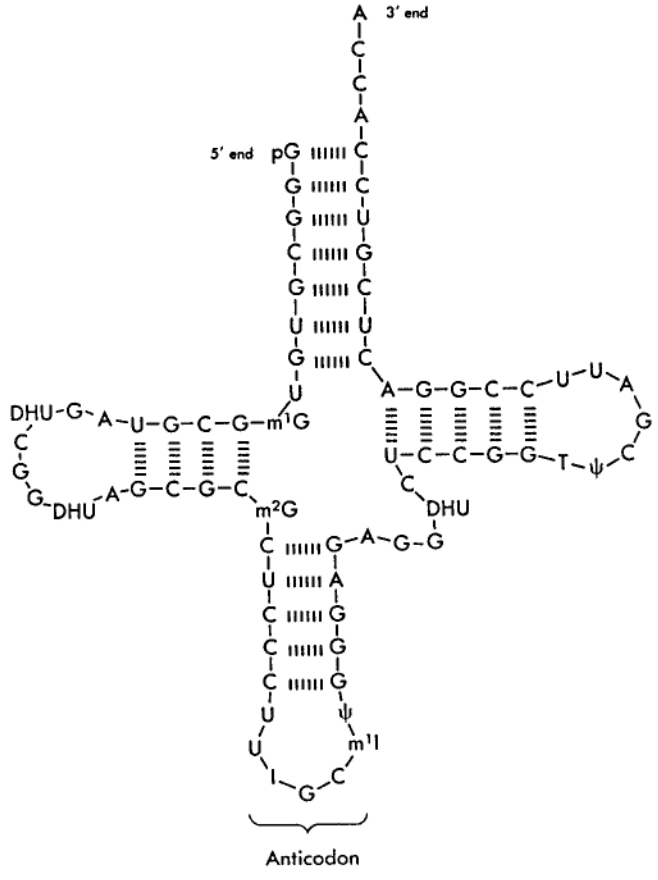


Figure 3-14
Yeast alanine tRNA structure, as determined by R. W. Holley and his associates. The anticodon in this tRNA recognizes the codon for alanine in the mRNA. Several modified nucleosides exist in the structure: ψ = pseudouridine, T = ribothymidine, DHU = 5,6-dihydrouridine, I = inosine, m^1G = 1-methylguanosine, m^1I = 1-methylinosine, and m^2G = *N,N*-dimethylguanosine.

From J. D. Watson et al
Molecular Biology of the Gene.

Figure 5

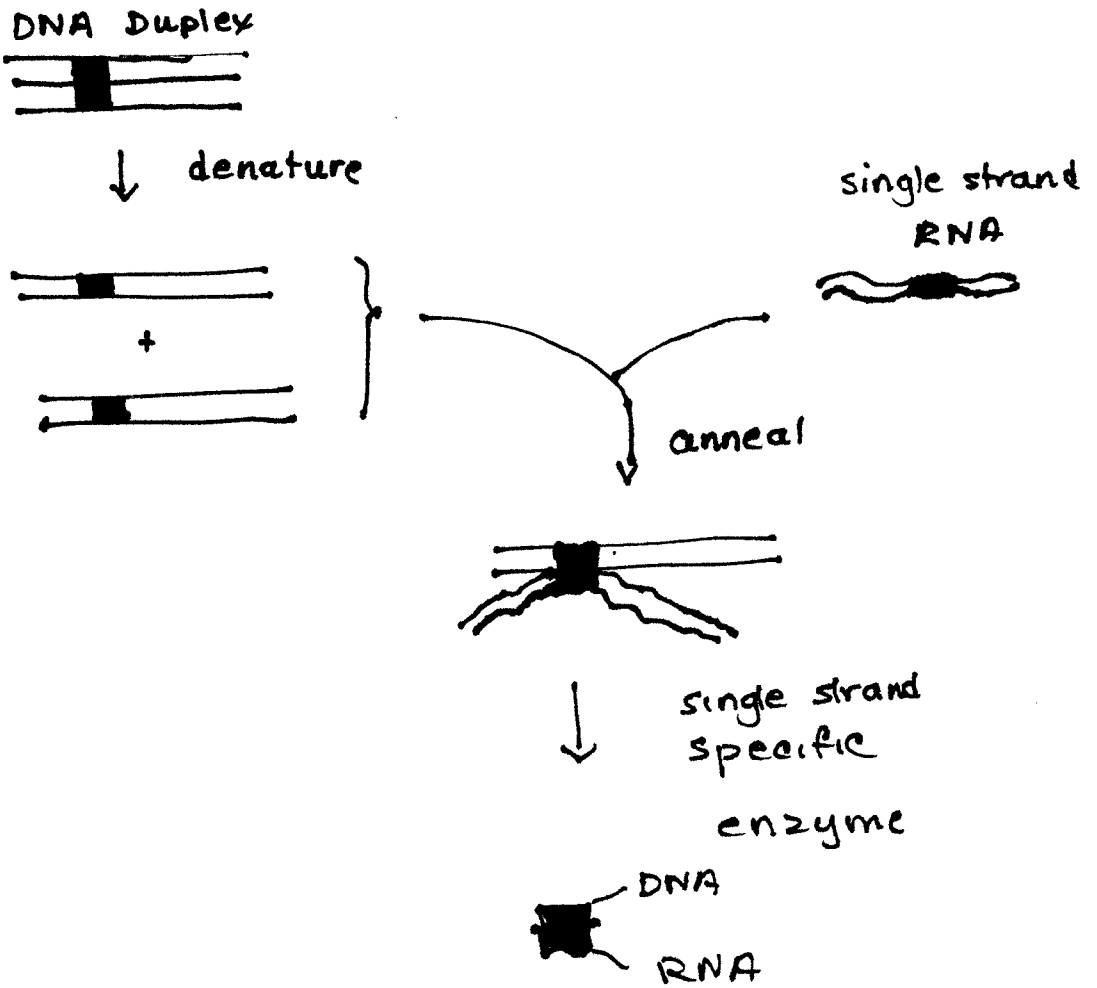

DNA Duplex
denature
+
single strand
RNA
anneal
single strand
specific
enzyme
DNA
RNA

Figure 6

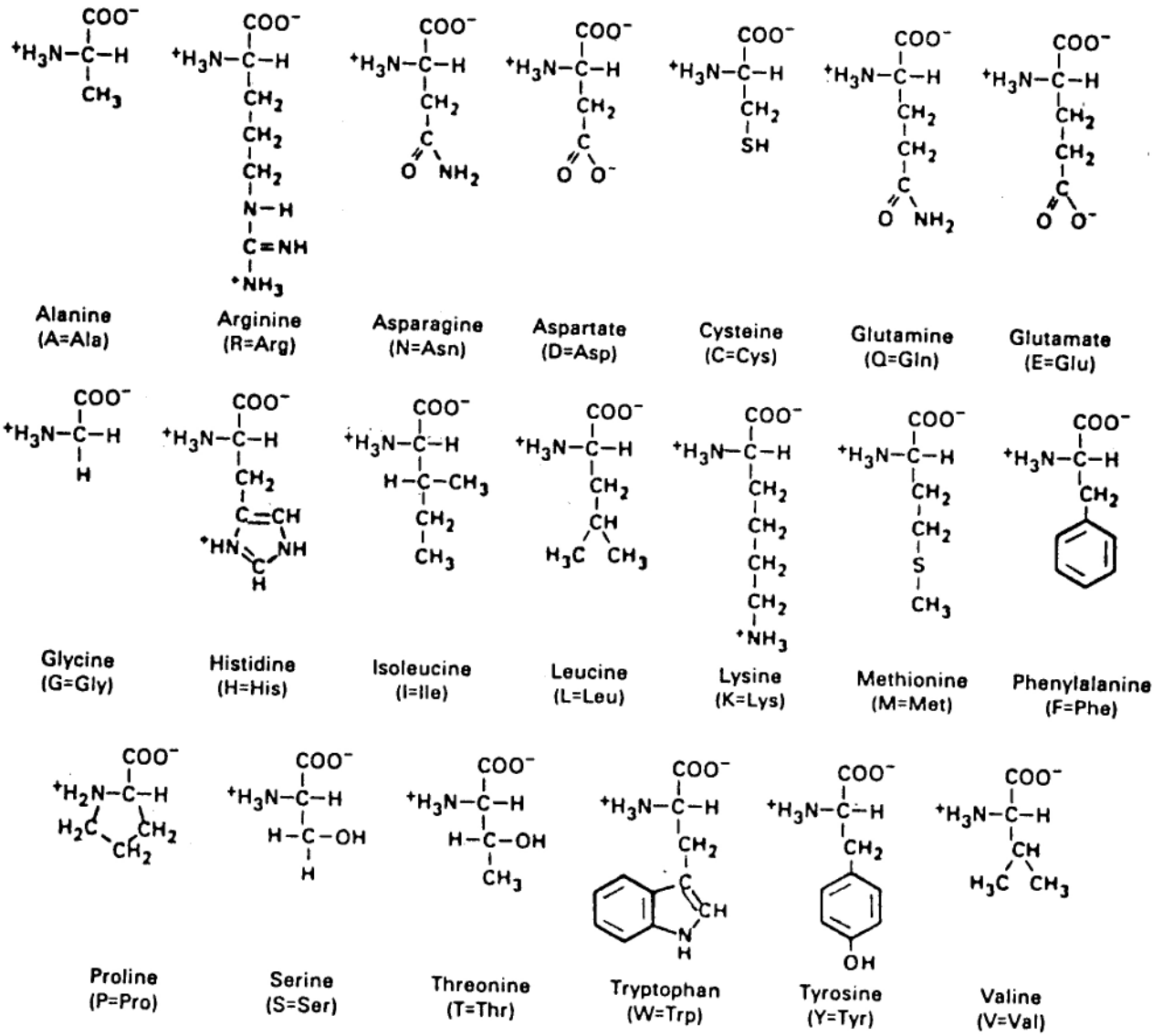

Alanine (A=Ala)
Arginine (R=Arg)
Asparagine (N=Asn)
Aspartate (D=Asp)
Cysteine (C=Cys)
Glutamine (Q=Gln)
Glutamate (E=Glu)
Glycine (G=Gly)
Histidine (H=His)
Isoleucine (I=Ile)
Leucine (L=Leu)
Lysine (K=Lys)
Methionine (M=Met)
Phenylalanine (F=Phe)
Proline (P=Pro)
Serine (S=Ser)
Threonine (T=Thr)
Tryptophan (W=Trp)
Tyrosine (Y=Tyr)
Valine (V=Val)

Figure 7

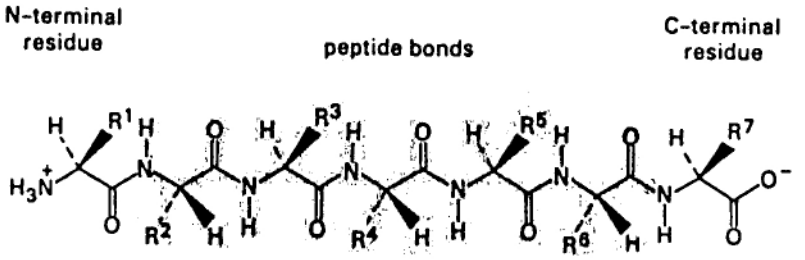


Figure 8

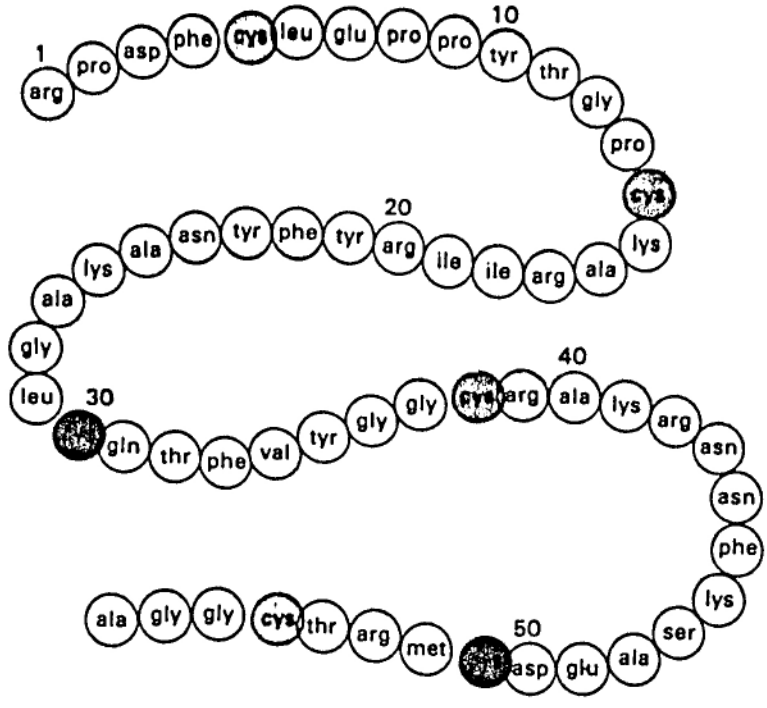

1
arg
pro
asp
phe
cys
leu
glu
pro
pro
10
tyr
thr
gly
pro
cys
lys
20
arg
ile
ile
arg
ala
asn
tyr
phe
tyr
ala
lys
ala
gly
leu
30
gln
thr
phe
val
tyr
gly
gly
cys
arg
40
ala
lys
arg
asn
asn
phe
lys
ser
ala
glu
asp
50
met
arg
thr
cys
gly
gly
ala

Figure [redacted] 9

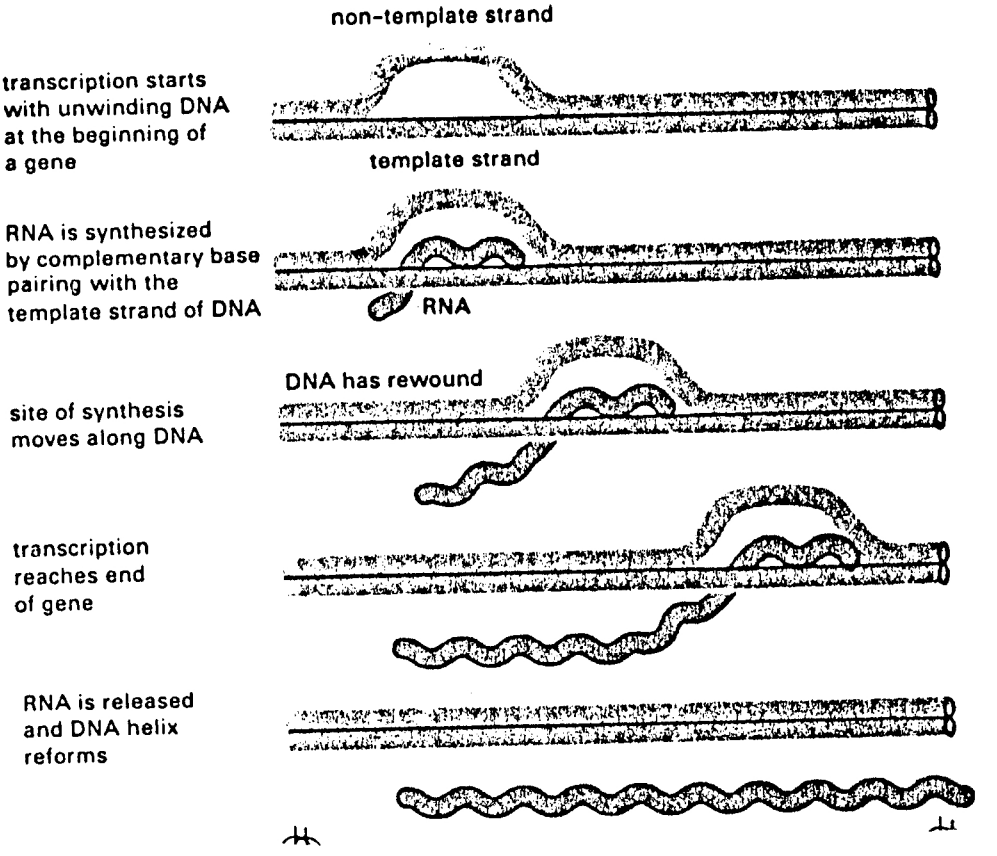


Figure 10

Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ileu	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
GCA	AGA	GAT	AAT	TGT	GAA	CAA	GGA	CAT	ATA	TTA	AAA	ATG	TTT	CCA	AGT	ACA	TGG	TAT	GTA	TAA
GCG	AGG						GGG		ATT	TTG				CCG	AGC				GTG	TAG
GCT	CGA									CTA				CCT	TCA	ACG			GTT	TGA
GCC	CGT						GGT		ATC	CTG			TTC	CCC	TCG	ACT		TAC	GTC	
	CGC						GGC			CTT					TCT	ACC				
	CGG									CTC					TCC					

Figure [redacted] 11

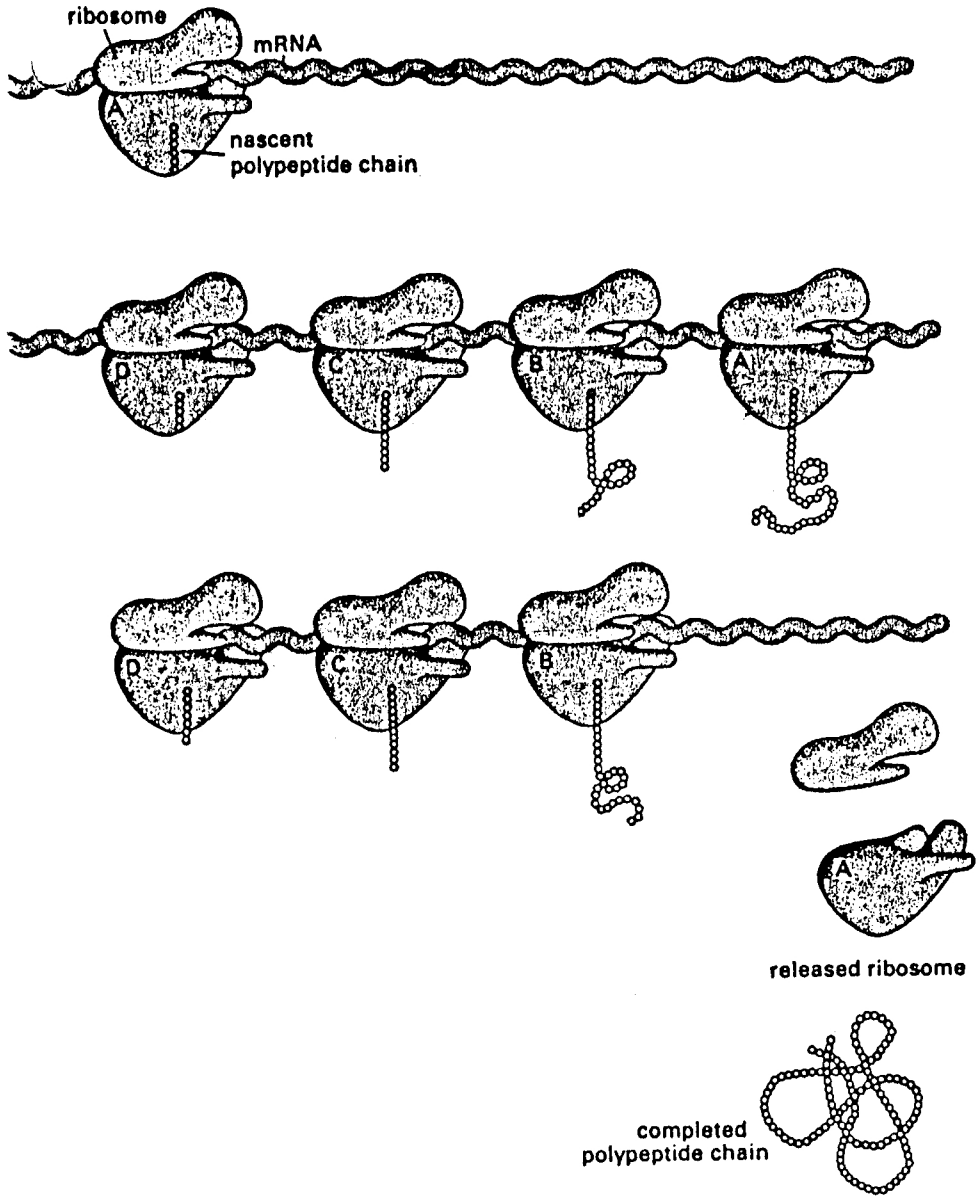


Figure 12

messenger RNA

5'- AUGUUUGAAGGAACAUAA- 3'

NH_2- met·phe·glu·gly·thr·- COOH

polypeptide

12A

Figure

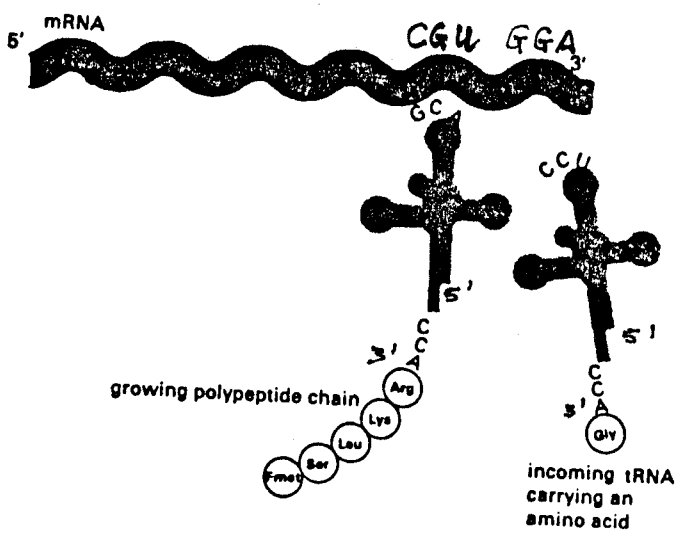

mRNA
5'
CGU GGA
3'
GC
CCU
5'
5'
C
C
A
3'
growing polypeptide chain
Arg
Lys
Leu
Ser
Fmet
5'
C
C
A
3'
Gly
incoming tRNA
carrying an
amino acid

Figure 13

Reading frame

5' 1 2 3
CAGUCUAUGGCAAAUAAGGUAGACCAU- 3'

1 gln · ser. met. ala. asn. lys. val. asp. his

2 ser. leu · trp. gln. ileu. arg. STOP

3 val. tyr. gly. lys. STOP

14

Figure ~~[illegible]~~

Splicing of a primary transcript removes intron sequences yielding mature messenger RNA

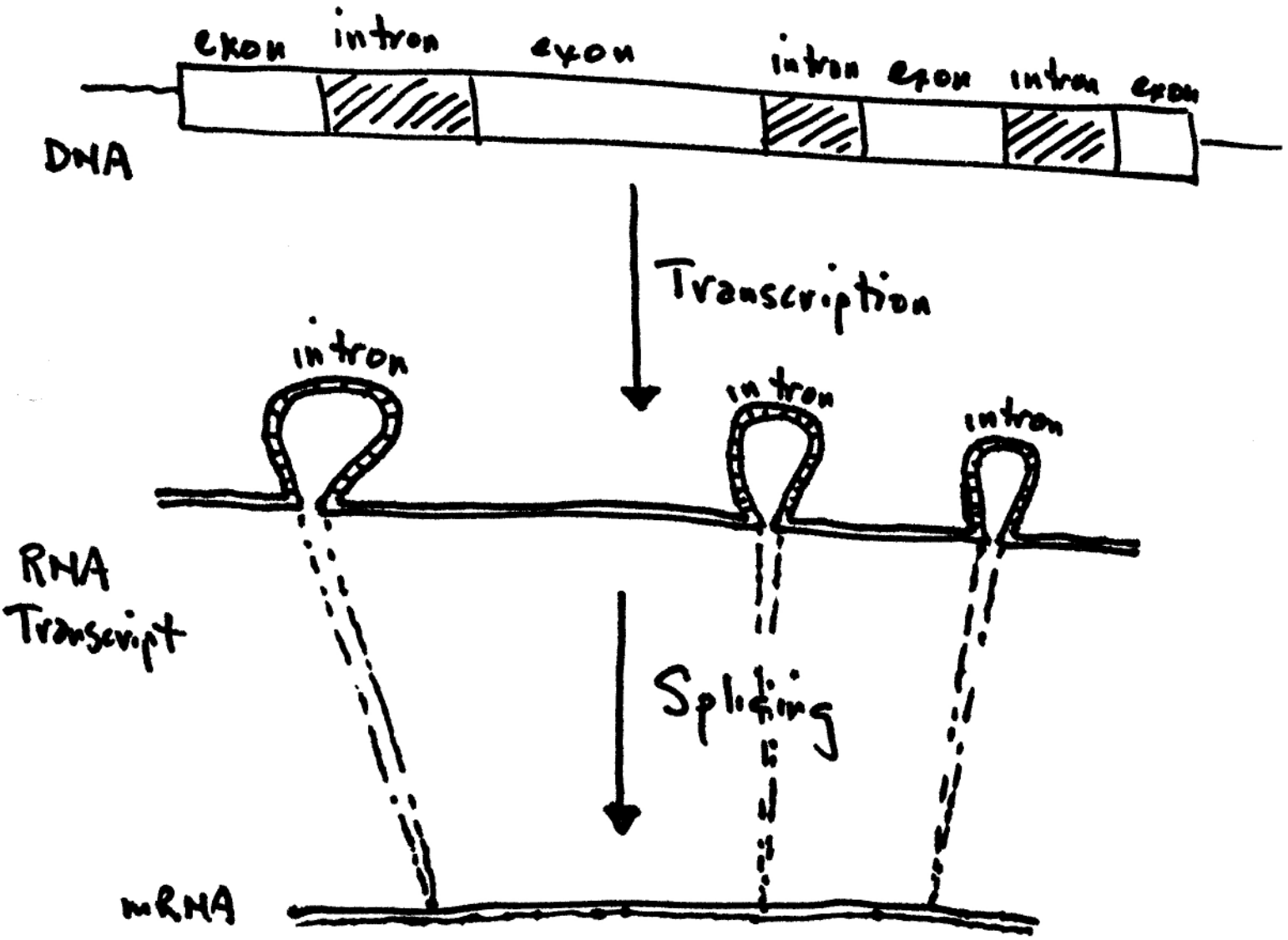


Figure ~~15~~

Alternative Splicing

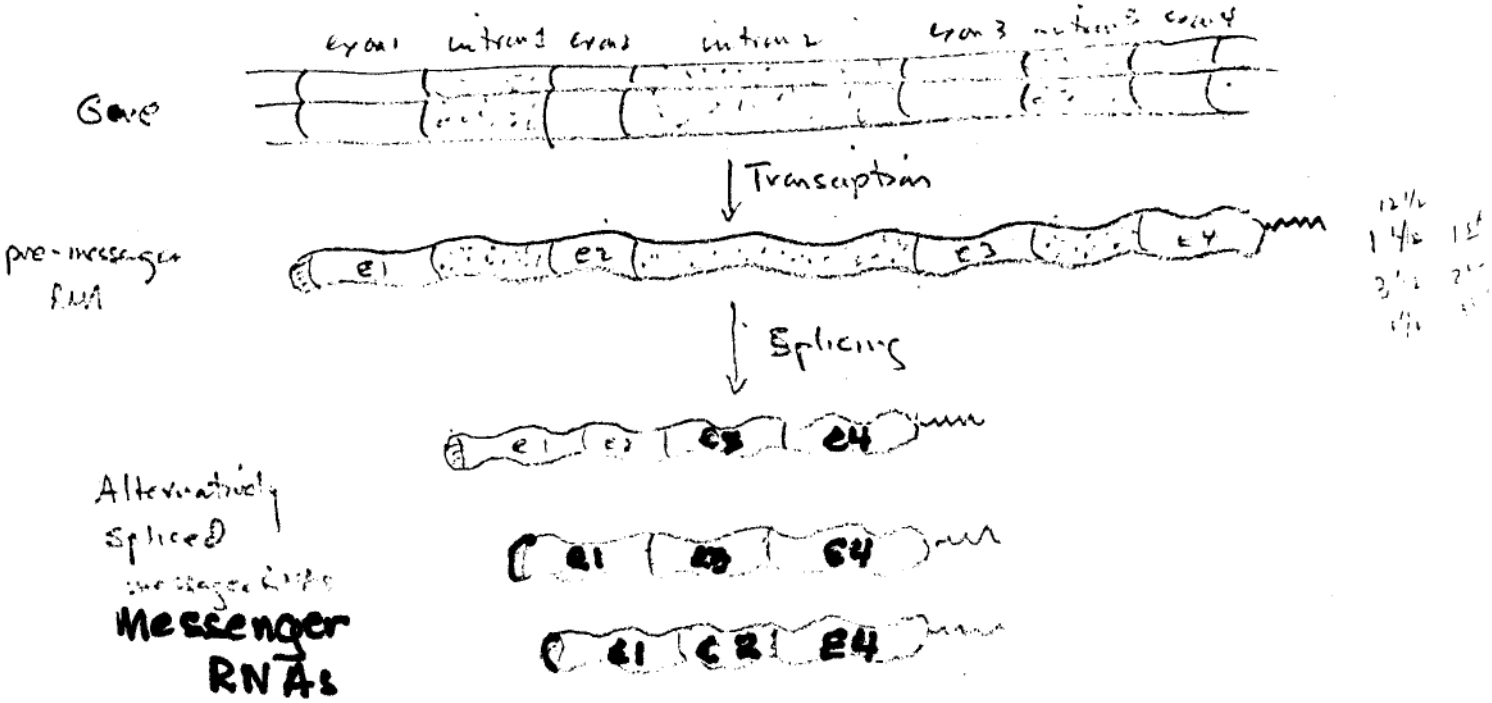
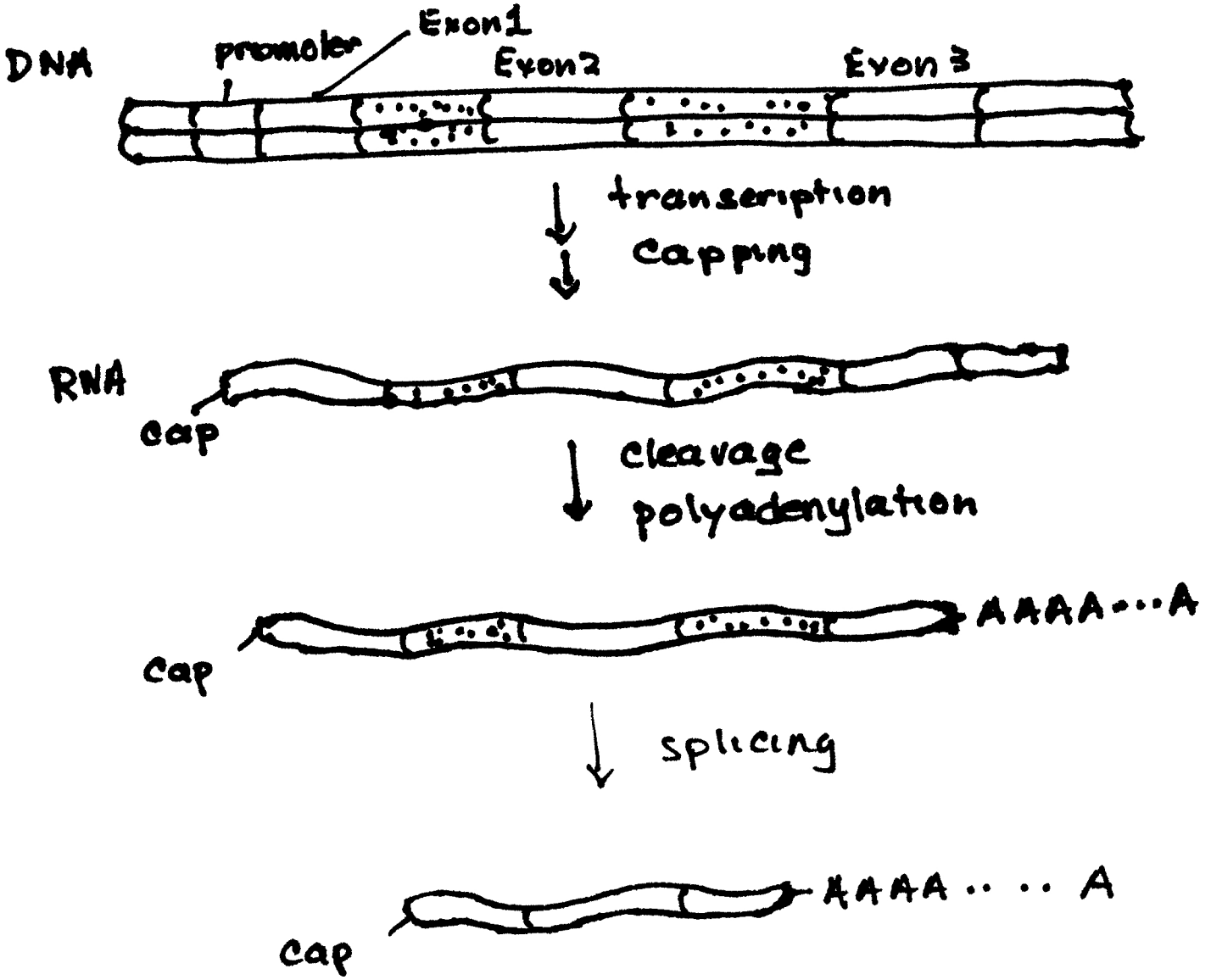
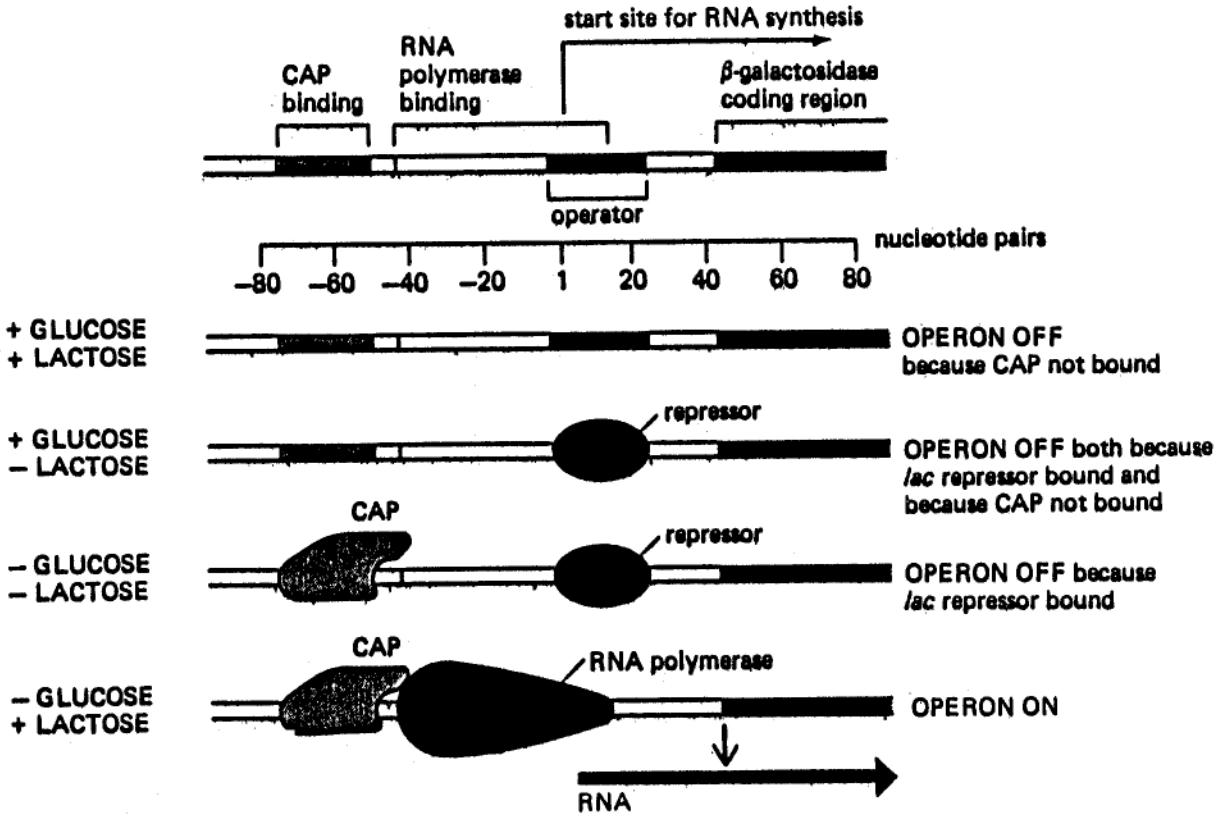


Figure 16



cap = 7- methylguanosine - PPP-

Figure 17



from B. Alberts et al
Molecular Biology
of the Cell.