

BASIC CONCEPTS OF MOLECULAR BIOLOGY

Randy J. Horwitz, MD, PhD

MOLECULAR BIOLOGY: AN OVERVIEW

The field of molecular biology has grown exponentially during the past decade. Nowhere else in medicine is the marriage of basic and clinical science more apparent than in this rapidly changing field of study. Today's clinician can no longer ignore the advances in the field of molecular biology that are affecting every specialty and subspecialty of medicine. Diagnostic testing has evolved to the point that molecular defects are no longer detected at the protein level, but rather at the level of the gene.

The field of allergy and immunology has used the tools of molecular biology to a great extent. Diagnoses of immunodeficiencies have been refined such that defects in individual cellular signaling proteins (e.g., ZAP-70 kinase) have been implicated as causes of certain B-cell deficiencies. In addition, with the advent of recombinant allergen extracts, the field of immunotherapy benefits from the standardization of numerous reagents. Perhaps in the near future, all allergists will be using the standardized recombinant extracts for testing and therapy, and universal definitions of atopy will be accepted.

This article serves as an introduction and overview of some of the more common molecular biology techniques that may be encountered in practice as well as in the literature. To establish a common starting point, this article describes those biologic processes central to the under-

From the Section of Allergy and Immunology, Department of Medicine, University of Wisconsin Hospitals and Clinics, Madison, Wisconsin

standing of these techniques. For readers seeking additional background information, several excellent textbooks of molecular biology are available.^{1, 9, 14, 18}

SUMMARY OF MOLECULAR BIOLOGY PROCESSES

It is impossible to adequately examine all aspects of molecular biology in one brief article, but some review is essential in order to understand the commonly used techniques discussed later in this article. The analogies used are not meant to oversimplify, but rather are an attempt to clarify complex concepts.

The central dogma of molecular biology is illustrated in Figure 1. The flow of genetic information ultimately results in the production of proteins, both the ubiquitous housekeeping proteins as well as the cell-specific molecules. Housekeeping proteins are required by all cells to sustain basic life support (or maintenance) functions. Examples of such proteins are actin or glyceraldehyde-3-phosphate dehydrogenase (G3PDH). Cell-specific proteins are those exclusively used by only certain cell types, such as insulin produced by pancreatic cells, or immunoglobulins expressed by B cells. In the 1970s, David Baltimore and Howard Temin independently described reverse transcriptase, a viral protein capable of making DNA out of RNA.³ They subsequently shared the Nobel Prize for their discovery. This is the only major modification to the dogma that has occurred thus far. From Figure 1, one may deduce that one major function of DNA is to provide a template for protein synthesis.

A conceptual approach to this central paradigm is to consider the cell equivalent to a building construction site (Fig. 2). Most of the brainwork occurs in the cell's nucleus, which is much like the construc-

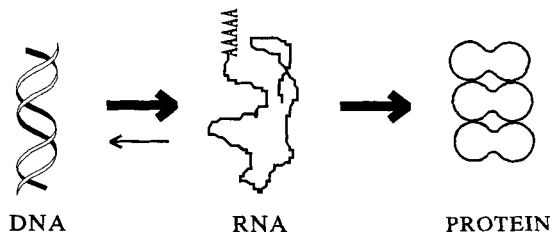


Figure 1. The central dogma of molecular biology. The flow of genetic information in the cell generally follows this scheme. Genetic information is stored as DNA, a stable molecule capable of self-replication. When needed, certain portions of the DNA molecule are transcribed into RNA, a molecule that carries the information from the nucleus to the cytoplasm. RNA may in certain situations be replicated or even converted back to DNA. Its usual fate, however, is to be translated into a functional protein molecule, the ultimate step in gene expression.

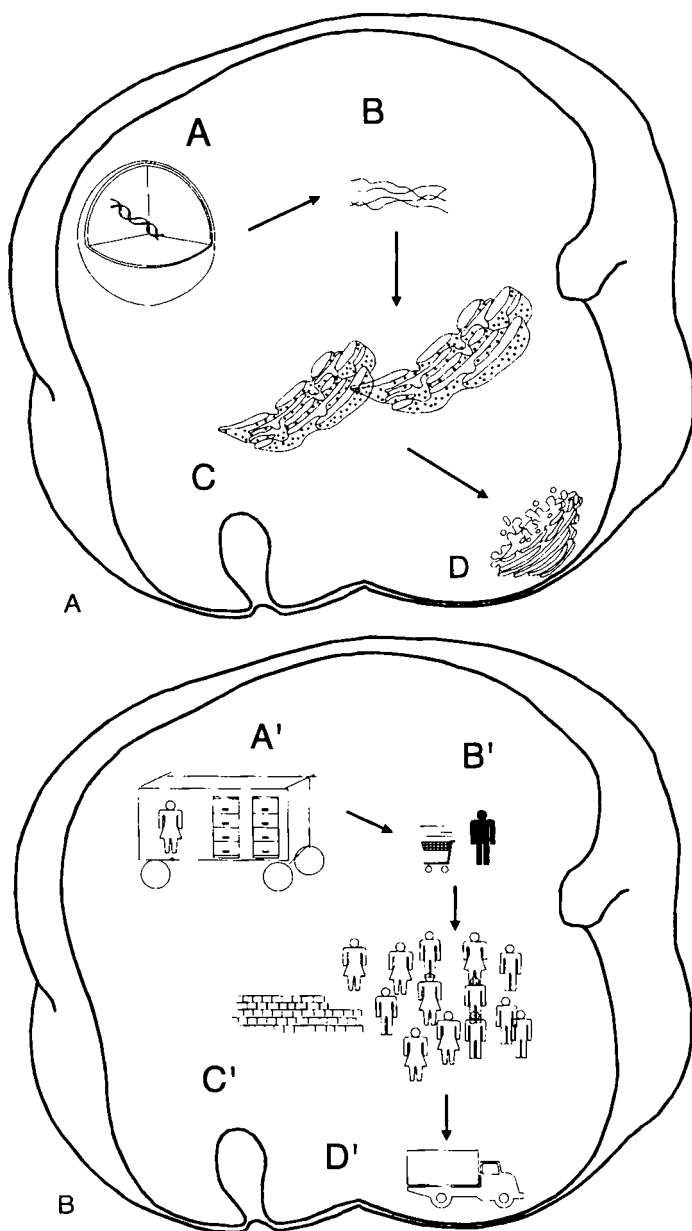


Figure 2. Molecular biologic processes in the cell. Gene expression in the eukaryotic cell (A) is similar to activities at a construction site (B). **A, (A')**: Information is stored as DNA (the blueprints) in the nucleus (the trailer). The RNA polymerase (the architect) decides which parts to transcribe (copy). **B, (B')**: An mRNA molecule (the messenger) delivers the selected plans to the cytoplasm (work site). **C, (C')**: Once in the cytoplasm, the ribosomes (construction workers) begin using amino acids (bricks) to assemble the functional protein molecule (the building). **D, (D')**: The finished protein either stays in the cell as a structural component or metabolic enzyme or may be packaged and exported via the golgi apparatus. Similarly, parts of the construction project may be shipped to other work sites for modifications. Details can be found within the text.

tion company trailer. Within the trailer/nucleus can be found the detailed plans for the entire building (or organism), the architect's blueprint (the DNA), as well as the architect (the RNA polymerase). RNA polymerase, the architect in this scenario, scans the DNA, determines which portions (genes) need to be transcribed, and begins the production of the mRNA molecule. The polymerase does have some help in identifying what regions of the DNA are important to transcribe into RNA, in the form of transcription factors. These molecules act like little flags, attracting the polymerase molecule and identifying regions that are important to transcribe. In general, each portion of DNA that is thus copied into RNA forms a unit known as a gene. The polymerase uses the master DNA molecule as a template from which to synthesize the RNA. Accuracy in copying (or "transcribing") is paramount as any error at this stage subsequently will be manifest as an abnormal protein.

There is one important rule about the blueprint, however: it can never leave the trailer. The dilemma then is how to get the plans out to the construction workers at the actual worksite (the ribosomes in the endoplasmic reticulum). Nature has solved this problem by using a specialized messenger molecule called messenger RNA (mRNA). This is the gopher of the construction site, responsible for copying those pertinent portions that are needed at that time (or in that particular cell), and then leaving the trailer (nucleus) and running out to the worksite (the endoplasmic reticulum) and hand-delivering them to the workers (the ribosomes). Here, the plans carried by the gopher (mRNA) dictate the assembly of that particular part of the building (a protein). The assembly of the protein is carried out by bricklayers (transfer RNA molecules; tRNA) using high quality bricks (amino acids) that together make the building (the completed protein).

In reality, DNA is the master blueprint in which almost every characteristic in an organism is described and archived. Nearly every cell in the body contains a complete copy of an individual's DNA. Information specifying (or "coding for") everything from eye color to the number of fingers on each hand is contained within each cell—even those cells uninvolved in those features. Clearly some mechanism must exist to selectively use (or "express") only the information that is essential to that particular cell. For example, there is no need for a liver cell to use the gene responsible for eye color. Every gene is represented in the DNA, but only certain genes are expressed. This regulatory role has been largely assumed by the RNA molecule.

DNA: STRUCTURE AND PROPERTIES

An appreciation of the history of molecular biology usually begins and ends with the now famous proposal of the structure of DNA by Watson and Crick in 1953.²⁰ Needless to say, there is a vast and rich

history of scientific discovery both preceding and following this notable accomplishment.

Deoxyribonucleic acid was discovered in the late 1860s by Friedrich Miescher, a Swiss chemist who was examining cell nuclei.¹¹ At the time, its function was unknown. In fact, even after Gregor Mendel's classic observations on the inheritance of morphologic traits in pea plants gained acceptance in the 1900s, no one had associated his basic unit of inheritance, the "gene," with DNA. In 1928, Frederick Griffith demonstrated the ability to change a phenotypic characteristic of a recipient bacterial strain by transferring the characteristic from a dead donor colony, a process known as *transformation*.¹⁰ The implication of his work was that a nonliving molecule or factor is capable of mediating such a transfer of genetic characteristics. The identity of the "transforming" molecule remained a mystery until 1944, when three workers, Avery, MacLeod, and McCarty, identified these transforming molecules as DNA.² Using enzymes that preferentially degrade DNA, RNA, or proteins (DNase, RNase, and protease, respectively), they demonstrated that only DNase had an effect on the transformation of the bacteria.

Watson and Crick were aided in their delineation of the structure of DNA by several pieces of evidence. Edwin Chargaff previously had shown a relationship between the relative ratios of the nucleotide bases present in DNA samples from numerous sources.⁷ He found that the concentration of adenine (A) in the samples always equaled the concentration of thymidine (T). Similarly, the number of guanine bases (G) was equal to that of cytosine (C). Watson and Crick also had access to X-ray diffraction patterns of isolated DNA molecules that were prepared by R. Franklin and M. Wilkins. All these data were used in the formulation of the double-helix structure of the DNA molecule.

The familiar double helix resembles a spiral staircase, with the railings being the phosphodiester-carbohydrate backbone, and the steps representing the nucleotide bases (Fig. 3A). Hydrogen bonds between complementary bases hold the two strands together. There are only four bases in DNA: adenine, guanine, cytosine, and thymidine (commonly abbreviated A, G, C, and T, respectively) (Fig. 3B). It is the order of these bases within the DNA molecule that ultimately dictates the order of the amino acids in the protein that is formed (see section on protein synthesis). Thus, using what is essentially a four-letter alphabet, the DNA molecule is able to store enough information to describe and synthesize every protein found in an organism. Using this four-base system, it would require a book of 500,000 single-spaced pages to list all of the bases of the human genome. Fortunately, nature is able to save space by winding and coiling the DNA around specialized structures and cramming it into the cell's nucleus for safe-keeping.

Each of the two strands comprising the DNA molecule is complementary, rather than identical. That is, whenever cytosine occurs in one strand, it will be paired with guanine in the other strand. Similarly, whenever thymidine occurs in one strand, it will be linked to its complementary nucleotide, adenine. This provides the explanation for

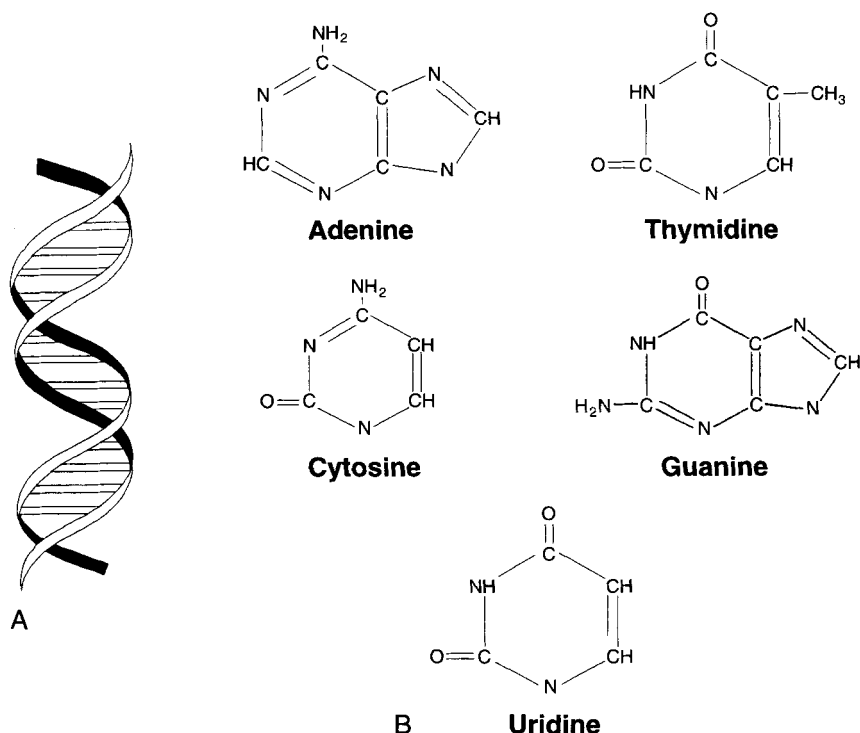


Figure 3. The structure of DNA. The DNA molecule consists of two strands of nucleotide polymers held together by complementary hydrogen bonding between nitrogenous bases. It is arranged in a double helix configuration, not unlike a spiral staircase (A). The nucleotides are composed of a carbohydrate (deoxyribose in DNA; ribose in RNA), a phosphate moiety, and one of four nitrogenous bases. DNA uses adenine, guanine, cytosine, and thymidine. RNA uses uracil in place of thymidine (B).

Chargaff's findings of base concentrations noted previously ($[A] = [T]$ and $[G] = [C]$). One DNA strand is, thus, considered the complement of the other strand, and carries the same genetic information. This fact is critical to the replication (copying) of the DNA molecule—with two copies (two strands) of each "blueprint" per DNA molecule, every time the strands separate and duplicate themselves, they generate an exact copy of the original molecule. More details of this replication are provided later, but first an explanation of how all that genetic information is put into a tiny molecule is in order.

Eukaryotic Chromosome Structure

Completely stretched out, the DNA from a human cell would measure almost 1 m long. The only way to fit this size molecule into a 10-

μm nucleus is to wrap it very tightly around itself (like a ball of twine), or around some type of spool (like a spool of thread). In eukaryotes, nature uses both of these motifs. First the DNA is spooled tightly around a protein structure called a *nucleosome*. The nucleosome, in turn, is composed of an aggregation (actually an octet) of four distinct types of nuclear proteins called *histones*. These histones are largely made up of basic amino acids, and are highly positively charged. This net positive charge helps explain why they bind so tightly to the negatively charged DNA molecule.

Each nucleosome has approximately two turns (or 146 base pairs) of the DNA molecule wound tightly around it, with only about 20 to 80 base pairs separating each nucleosome (Fig. 4). This arrangement of nucleosomes and DNA gives the appearance of closely spaced "beads-on-a-string" when visualized by electron microscopy. One might think that this coiling of the DNA would result in a structure compact enough to fit into the nucleus, but this is not the case. The next level of chromosomal structure used by the cell requires the nucleosomes to wrap around themselves, thus condensing the DNA even further. To accomplish this, a fifth type of histone molecule, called *histone H1*, is used.

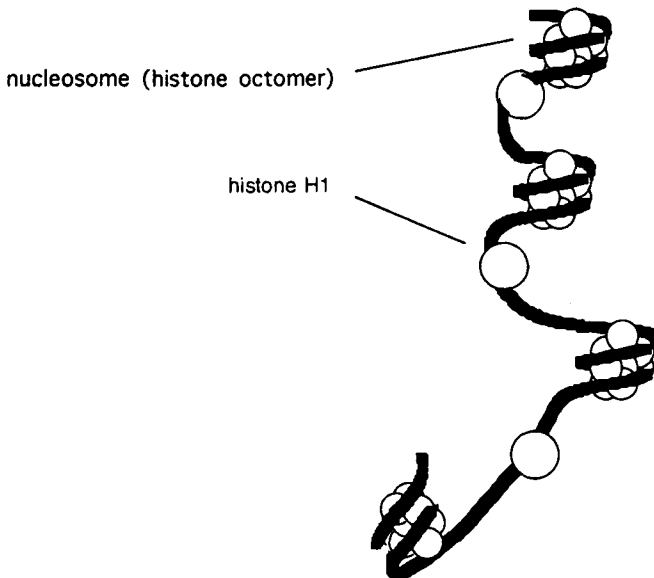


Figure 4. The structure of chromatin. DNA assumes a highly compact form when combined with nucleoproteins. This structure is called *chromatin* and is formed when the DNA strands, already in their double helix configuration, wrap around proteinaceous particles called *nucleosomes*. The nucleosomes are in turn composed of eight subunits, called *histones*. One type of histone, H1, is used to further condense the DNA, allowing the nucleosomes to wrap around themselves. The result of these higher-order structures is a complex, dense molecule that is easily able to fit into the nucleus.

Histone H1 is the only histone protein not associated with the nucleosome core. It is much larger than the other four histone proteins, and is found between the nucleosomes (in that 20–80 base pair linker region). The nucleosomes are packed together with the aid of histone H1 to form a thick fiber of approximately 30 nm in diameter. The process of DNA packaging does not stop here. Further wrapping and condensation ultimately form a super-compact nuclear DNA molecule called *chromatin*. The only form of DNA more compact than chromatin is the metaphase chromosome, a structure some 8000 times more compact than the original DNA strand.

DNA Replication

One of the amazing features of DNA is how it replicates. The Watson-Crick model for DNA structure was unique in that it predicted a means of replication based upon the presence of two complementary strands.^{19, 20} The actual process of replicating the entire genome (all the DNA sequences in an organism) is straightforward in principle, but reveals a level of complexity that has evolved through generations of lower organisms. In a nutshell, the double helix separates into two strands in an action resembling the unzipping of a zipper. A new pair of daughter strands is then generated through the action of DNA polymerase. Each of the two parental strands serves as a template to guide the synthesis of the newly formed complementary strand. Thus, what begins as one DNA molecule composed of two complementary strands replicates to yield a set of two DNA molecules, each composed of one “new” strand and one “old” strand. The process whereby one original (or parental) strand is used as a template for the synthesis of its new partner strand is called *semi-conservative replication* (Fig. 5).

Most studies examining the mechanism of DNA replication have been performed in prokaryotes. This is because of the relative ease in obtaining bacterial isolates deficient in one or more replication enzymes, compared with eukaryotic cell lines. Despite this fact, it appears that the basic mechanisms of DNA replication are similar in both systems, with few differences.

The process of DNA replication in prokaryotes proceeds at a rate of approximately 500 nucleotides per second. There is a single origin of replication defined in bacteria, which represents the start (and eventually, the stop) site of DNA replication. In addition, the DNA in prokaryotes is “naked,” or free of the nucleosome/histone protein spools seen in eukaryotic nuclear DNA. This difference is manifested in the relative rates of replication. In eukaryotes, because the DNA is packaged in such compact nucleosome-bound arrangements, the rate of replication is only 50 nucleotides per second. Using this rate of replication, it would require almost 800 hours to replicate an average human chromosome. To compensate for this 10-fold slower replication rate, eukaryotes utilize multi-

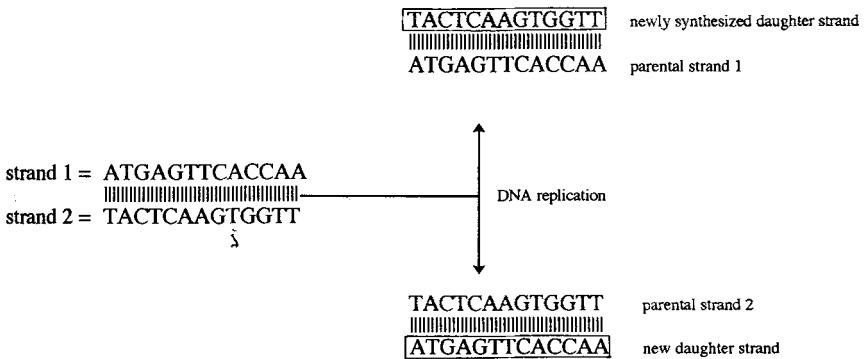


Figure 5. DNA replication. The DNA molecule replicates by a process known as *semiconservative replication*. That is, two “parental” strands separate and each separate strand then becomes the template for new “daughter” strands. Complementary base pairing with the template determines the order of nucleotides added to the newly forming strand. The replication is semiconservative because each new double-stranded DNA molecule is composed of one old and one new strand.

ple replication origins, occurring every 30,000 to 300,000 base pairs throughout the genome.

Prior to summarizing the molecular processes involved in DNA replication, one crucial structural detail must be mentioned. As stated previously, DNA is normally found as a double helix, with each strand complementary to its partner. This arrangement is only possible because the strands demonstrate polarity (a difference between the “top” of the molecule and the “bottom”). At the 5' end of the DNA strand is the phosphate group of one terminal nucleotide, and at the 3' end of the strand is the hydroxyl group of the other terminal nucleotide. The complementary strand is also polar, but runs anti-parallel to its partner (i.e., one strand runs 5' → 3', and the other runs 3' → 5'). The nucleotides are linked together by covalent phosphodiester bonds between the 5'-carbon of one deoxyribose group and the 3'-carbon of the next. This feature of DNA structure is important because during DNA replication, new strands are only synthesized in the 5' → 3' direction. Each new nucleotide is added onto the 3'-hydroxyl group of the growing daughter strand. This has important implications in terms of the replication of each strand.

Regions where DNA replication begins are known as *replication forks*, because of their appearance under the electron microscope. In eukaryotes, many forks are present, with DNA replication proceeding bidirectionally from each one, until all the “growing bubbles” meet, and the entire genome has been replicated. The process of DNA replication involves numerous proteins interacting with the DNA at these replication forks (this is the case in both prokaryotes and eukaryotes). Two important proteins, *DNA helicase* and *single-strand DNA binding protein*

work in concert to help destabilize and unwind the helix to allow replication to occur. They both coat the DNA strands, unwind the DNA, and sort of straighten out the helix to permit efficient copying by the DNA polymerase.

In eukaryotes, at least two different polymerases are used to copy DNA. On what is called the *leading strand*, DNA polymerase δ is used to make daughter strands. Because the parental strand is being unwound in the $3' \rightarrow 5'$ orientation (opposite to the direction of synthesis), continuous polymerization is possible, and building a new daughter strand is straightforward. On the *lagging strand*, however, it is a different story. On this strand, the new daughter strand must be synthesized in a discontinuous fashion, restarting every 200 bases, using small RNA fragments acting as primers (Fig. 6). These RNA primers are made by an enzyme called *DNA primase*, and eventually are removed and filled in with DNA by a special DNA repair system. A different polymerase molecule, DNA polymerase α , is used to replicate DNA on the lagging strand.

DNA polymerase δ has what is termed a “proofreading” activity, meaning that it possesses a nuclease activity capable of removing an incorrectly inserted nucleotide. Polymerase δ checks each nucleotide for correct base pairing with the template as it is being added, and replaces the bad nucleotide instantly. Any incorrect nucleotides that the polymerase misses (about 1 per 10,000–100,000 bases) is corrected by one of the

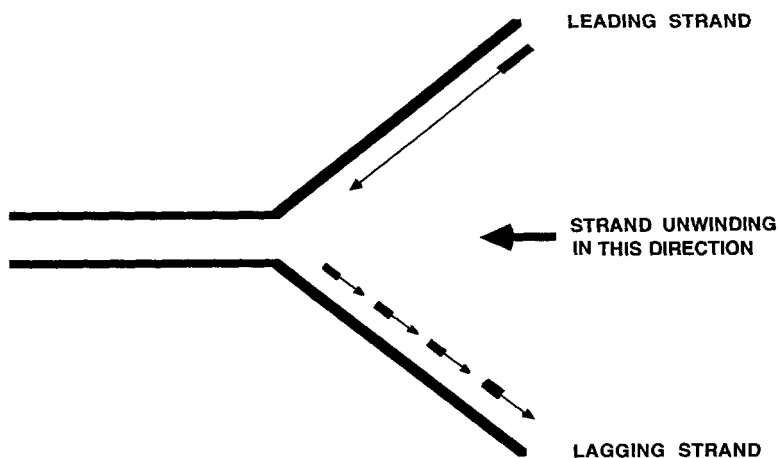


Figure 6. Discontinuous replication of the DNA molecule. The DNA polymerase used to replicate DNA can synthesize new molecules only in a $5' \rightarrow 3'$ direction. Because the DNA strands are arranged in an anti-parallel orientation, one of the strands must be replicated discontinuously. This is accomplished by creating small RNA primers that serve as starting points for replication. On the discontinuous or lagging strand, multiple primers must be repeatedly placed along the length of the strand to copy all of the sequence information. Such a problem does not arise on the leading or continuous strand.

two other eukaryotic polymerases, β and ϵ , whose sole purpose is to repair mutated or damaged DNA in the cell. Because DNA polymerase α , the lagging strand polymerase, has no proofreading function, these two repair enzymes help monitor the fidelity of replication on that strand as well.

RNA: STRUCTURE AND PROPERTIES

RNA differs structurally from DNA in several respects, which are summarized in Table 1. Thymidine, one of the four nucleotide bases present in the DNA molecule, is replaced with uracil in RNA. These two bases are closely related, and both bind well to the complementary base, adenine. The pattern of complementary base pairing in RNA is thus: A–U and G–C. This is important in the transcription process described later. The RNA polymerase reads the DNA sequence and builds the transcript by adding complementary ribonucleotides to the growing RNA molecule.

Another important difference between DNA and RNA is the secondary structure of the molecule. Whereas DNA exists as the familiar double-stranded helix, RNA is typically a single-stranded molecule. This does not preclude the formation of double-stranded regions within an RNA molecule, however. Local regions of short, complementary base pairing can occur at different sites within an RNA molecule, yielding configurations that resemble “stems and loops” with extensive secondary structure. This ability to form intramolecular bonds is important for the removal of noncoding regions of RNA (introns) following transcription (see section on transcription).

A final structural difference between DNA and RNA concerns the specific carbohydrate used to form the backbone of the molecule. DNA uses the sugar deoxyribose, whereas RNA contains ribose.

Table 1. COMPARISON OF RNA AND DNA MOLECULES

	Function	Structure	Sugar Used in Backbone	Cellular Location	Types of Molecules
DNA	Storage of genetic information	Double-stranded helix	Deoxyribose	Nucleus Organelles	—
RNA	Transmission of genetic information	Single-stranded molecule	Ribose	Nucleus Cytoplasm Organelles	hnRNA mRNA tRNA rRNA

Types of RNA

There are at least three important types of RNA molecules that have evolved to assume specific functions in eukaryotes. Messenger RNA (mRNA), as previously discussed, is produced via the transcription of specific regions of the DNA template in the nucleus. Among the RNA molecules of the cell, mRNA is the least abundant and the most heterogeneous in size, ranging from 200 base pairs up to approximately 3000 base pairs in length. The half-life of mRNA also varies considerably, a fact that has important implications for the regulation of gene expression. Despite this heterogeneity, there are several structural motifs that are conserved among all eukaryotic mRNA molecules. The mRNA cap is a special methylated nucleotide added to the upstream (5') end of the mRNA molecule following transcription. It is believed to inhibit rapid nuclease digestion of the mRNA, as well as to promote mRNA-ribosomal binding, thus facilitating translation. The 3' end of mRNA molecules commonly contain a poly-A tail, a series of 30 to 250 adenine nucleotides added by the action of a special poly-A polymerase following transcription. This structure, too, is thought to inhibit nucleolytic degradation of the mRNA molecule. There are several exceptions to the cap/poly-A tail motif—several types of histone mRNAs lack these structures. These mRNAs, however, are regulated by rather unique and complex mechanisms. Most mRNAs also contain long untranslated regions at the 3', 5', or both ends of the molecule. Although these sequences do not code for structural proteins, they are essential regulatory regions for control of gene expression and contain binding sites for numerous regulatory factors used in translation.

A second type of RNA molecule is the ribosomal RNA molecule (rRNA). This is the most abundant of the RNA species found in eukaryotic cells. The rRNA molecule is an essential component of the ribosome, the actual site of protein synthesis in the cytoplasm. The eukaryotic ribosome is composed of two heterogeneous subunits: a large and a small subunit. Each subunit consists of protein and rRNA in about equal proportions. There are several sizes of rRNA molecules along with nearly 100 different proteins in each eukaryotic ribosome. Because of the overwhelming requirement for protein synthesis in eukaryotes, rRNA molecules are found in very high concentrations in the cell, and the genes coding for rRNAs are present in multiple (often hundreds of) copies.

The third type of RNA found in the eukaryotic cell is the transfer RNA molecule (tRNA). This RNA, with 75 to 100 nucleotides, represents about 15% of the total RNA in a mammalian cell, and is exclusively found in the cytoplasm. The role of tRNA is to bring the amino acid to the ribosomes where translation occurs. During the process of protein synthesis, two tRNA molecules attach to the ribosome-mRNA complex at adjacent sites while peptide bonds form between the amino acids they are carrying. There are at least 20 species of tRNA in eukaryotic cells—one corresponding to each of the amino acids. Despite the large

number of different tRNAs, they all share similar structures and sizes. The molecule is shaped like an inverted cloverleaf, and has several important sites. The anticodon site may be likened to a key—only the correct anti-codon triplet binds to the codon of the mRNA template. This process assures that the correct amino acid is inserted onto the growing polypeptide chain. The other major “loops” in the tRNA molecule are the “D-loop” (containing dihydrouridine, an unusual nucleotide), and the “TΨC loop” (containing ribothymidine-pseudouridine-cytidine, three unusual nucleotides). These structures are thought to be important in tRNA-ribosome binding.

There are a few other minor RNA species found in eukaryotic cells, such as small nuclear RNAs (snRNA) that are involved in RNA splicing, and heterogeneous nuclear RNAs (hnRNA) that are the large primary transcripts found in the nucleus prior to splicing and secretion. These hnRNAs are the precursor molecules to the mature mRNAs, and are often difficult to isolate because of their transient existence. These molecules are discussed later.

Transcription

The action of copying selected areas of a DNA template to form an RNA molecule (DNA transcription) is one of the first steps in the conversion of a gene to a functional protein. As we shall see, it is also the key regulatory step in the regulation of eukaryotic gene expression. Transcription is mediated by the action of RNA polymerases, enzymes that elongate RNA chains by the addition of ribonucleotides onto the 3'-hydroxyl group of a growing chain. The polymerization reaction thus is similar to the action of DNA polymerases previously discussed. Unlike DNA polymerase, RNA polymerase has no proofreading activity, so it lacks the means to correct any misincorporations made during transcription. Because RNA is a transient molecule, and not a part of the permanent blueprint of the cell (and is not “passed on” during cell division like DNA), any minor error in transcription probably will not be harmful to the cell. In addition, because a great many transcripts of the same gene generally are made at a given time, a few bad apples (or faulty transcripts) will not spoil the whole bunch.

The transcription process differs in prokaryotes and eukaryotes in many respects, but two notable distinctions bear mentioning. As previously discussed, prokaryotic RNA lacks introns. Thus, with respect to transcription, “what you see is what you get,” or more specifically, everything in the DNA sequence of the gene will appear in the final transcript. The prokaryotic mRNA also lacks additional structures common to eukaryotic mRNA, such as the poly A tail and the 5' cap. Another important distinction is the “uncoupling” of transcription and translation. In prokaryotes, ribosomes begin binding to the RNA strand as it is being made. Thus, protein synthesis (translation) begins even before transcription has been completed. In fact, by the time the tran-

scription of the complete mRNA molecule is over, it has been coated with many ribosomes that are actively translating the protein. In eukaryotes, evolution has provided a nucleus that serves to compartmentalize the various processes of transcription and translation. The eukaryotic mRNA must be modified, processed, and finally exported to the cytoplasm prior to the initiation of translation.

There are three different RNA polymerase molecules used to transcribe eukaryotic nuclear DNA. They are imaginatively named RNA polymerases I, II, and III. It is RNA polymerase II that is responsible for the transcription of genes that eventually will be translated into proteins. RNA polymerase II is thus responsible for the production of mRNA. The other two polymerases transcribe RNA species that play a structural role in the process of protein synthesis. RNA polymerase I transcribes rRNA molecules, whereas polymerase III synthesizes tRNA molecules. We proceed to describe the mechanism of RNA polymerase II-mediated mRNA synthesis, and later expand on this subject in the discussion of the regulation of mRNA synthesis in eukaryotes.

At this point it is useful to explain the numbering of nucleotides. By convention, we write DNA and RNA sequences from left to right, in a $5' \rightarrow 3'$ orientation. Nucleotide numbering is centered on the first transcribed nucleotide, which is called number 1. Numbering continues positively to the right, and negatively to the left. Thus, the nucleic acid is represented like a number line, with both a positive and a negative orientation (except that there is no nucleotide numbered "zero"). The upstream direction in a gene is toward the $5'$ end, whereas the downstream is toward the $3'$ end of the sequence. In jargon, however, upstream sequences usually refer to the nontranscribed sequences found in front of the transcribed region, and downstream sequences are those located beyond the transcribed region of the gene.

There are basically three stages of transcription that we shall consider: initiation, elongation, and termination. The initiation phase is perhaps the most important, particularly when one considers that only about 1% of the total DNA sequence of a mammalian cell ever makes it out of the nucleus as mRNA. Certainly, some of the sequences are transcribed and then discarded as introns during RNA splicing, but the majority of the DNA sequences are not transcribed at all. Initiation begins when RNA polymerase II binds to the promoter region of the gene. A *promoter* is simply a region of the DNA, in front of (or upstream from) the transcribed part of a gene, to which the RNA polymerase binds. As the reader may recall, the beginning of many genes are "marked" by special sequences within the DNA. These sequences do not code for a protein—rather they are regulatory signals designed to tell the RNA polymerase important information about the gene, such as exactly where to begin transcription. The position of the promoter seems to be inflexible with regard to the transcription start site.

Many in vitro experiments using recombinant DNA have compared the transcription rates of genes before and after their promoter region was mutated or removed. In most cases, transcription of the genes with

deleted or mutated promoter regions was markedly reduced. Although every gene has its own distinct promoter regions, sequence analysis of eukaryotic genes identified several sequence motifs that nearly all genes have in common (so-called *consensus sequences*). One important sequence is called the *TATA box* (based upon its high thymidine and adenine content), or Hogness box (after its discoverer). This sequence of T-A-T-A is seen in the upstream (or 5') region of nearly all genes transcribed by RNA polymerase II. It is typically found 25 base pairs upstream of the transcription start site. The *CAAT box* is another sequence conserved in eukaryotic promoters. Its sequence is: GGCCAATCT. It, too, seems to be important in RNA polymerase binding and transcriptional start site recognition.

The first step in the transcription process occurs when an RNA polymerase molecule binds to the promoter region of a gene. The precise control of gene expression (i.e., deciding what genes will be transcribed and when) is facilitated by the presence of both proteins (called *transcription factors*) and specialized DNA sequences (called *enhancers*). The transcription factors bind to sequences within the promoter regions of specific genes and help the RNA polymerase bind tightly. The enhancer sequences are located at distant sites in the DNA and also serve to upregulate transcription. These are discussed in the following sections.

Following binding, the RNA polymerase molecule begins to break some of the hydrogen bonds between the DNA strands, thus separating (or melting) the two strands into a more open configuration. Approximately 50 nucleotides are opened in this fashion. One of the two DNA strands serves as the template for the RNA transcript. The choice of which strand to copy is largely a function of the location and orientation of the promoter region, as this sequence orients the RNA polymerase in one direction. There is not one particular strand that is used for all genes; it varies gene to gene. The RNA polymerase molecule continues to move along the DNA, melting as it moves, until it encounters the initiation site (nucleotide no. 1 in the nomenclature). At this point, it will incorporate the correct ribonucleotide triphosphate (from the vast, free-floating nuclear pool of nucleotides), and begin building the primary transcript. Phosphodiester bonds are formed between adjacent nucleotides in the chain.

The elongation process, like that of DNA replication, proceeds in a 5' → 3' direction. The RNA polymerase moves along the template in a step-wise fashion, unwinding the DNA just ahead of the growing primary transcript, thus exposing a new region for complementary base pairing. As the polymerase moves down the DNA template, the recently transcribed "melted" regions reassociate and reform the double-helix structure. This action also serves to displace the newly synthesized RNA molecule, so it is free to "dangle in the breeze," while growing even larger, until the elongation phase is completed. The unwinding and rewinding of the DNA molecule is performed by RNA polymerase, with some help from DNA unwinding proteins called *topoisomerases*.

The primary transcript continues to grow until a termination signal

is encountered by the RNA polymerase. In prokaryotes, it is the secondary structure of the growing transcript that facilitates its dissociation from the DNA template. Alternatively, a separate protein called "rho" binds to the nascent transcript and clips it off. In eukaryotes, however, less is known about the termination of transcription. It does appear that the primary transcript is cleaved at some point, but the polymerase molecule continues to ride along the DNA for a few thousand bases, making RNA transcripts that are destined for degradation.

The 5' end of the RNA molecule is the first part synthesized during transcription, and the specialized cap structure discussed previously is added to the transcript even before it is completed. In fact, cap structures can be isolated from transcripts that are only about 40 bases long. The poly-A tail, however, can only be added after the 3' end of the RNA transcript is completed. It is synthesized by the action of a separate polymerase called poly-A polymerase. This enzyme uses a consensus poly-A sequence to know exactly where to add the tail. Now we are left with a long primary transcript, complete with cap and tail, but it is not yet ready to be translated.

The average length of a primary RNA transcript (or hnRNA) is about 7000 bases, although molecules of up to 20,000 bases have been reported. The average size of a mature, cytoplasmic mRNA is probably closer to 2000 bases. It thus appears that some process is needed to "separate the wheat from the chaff," and get rid of all the excess RNA. This is achieved through a mechanism called *RNA processing* or *splicing* (Fig. 7).

RNA Processing (Splicing)

Prokaryotic RNA differs from eukaryotic RNA in several respects. Prokaryotic RNA is often polycistronic, meaning that each mRNA molecule may code for several proteins. Eukaryotic mRNA, on the other hand, is monocistronic, with each mRNA molecule translating one protein. Despite this fact, there are many regions of DNA transcribed into the monocistronic mRNA that do not code for part of the final protein. The discovery that many eukaryotic genes are "interrupted," made in 1977 by Louise Chow and others, was completely unexpected.^{4,8} Electron micrographs of mRNA molecules hybridized (bound) to their DNA templates revealed multiple large "loops" of DNA sequences that were not present in the mRNA molecule. The accepted dogma (up until that point) had predicted no extraneous DNA sequences because DNA is copied directly into mRNA. Suddenly, the absence of many long stretches of nucleotides in the mRNA molecule, mostly in the middle of the gene, had to be explained. As had been the case on numerous occasions, one remarkable series of experiments eventually led to a reappraisal and a rewriting of some revered rules of molecular biology.

These transcribed yet untranslated regions of DNA are called *intervening sequences*, or *introns* (from intervening regions). They are dis-

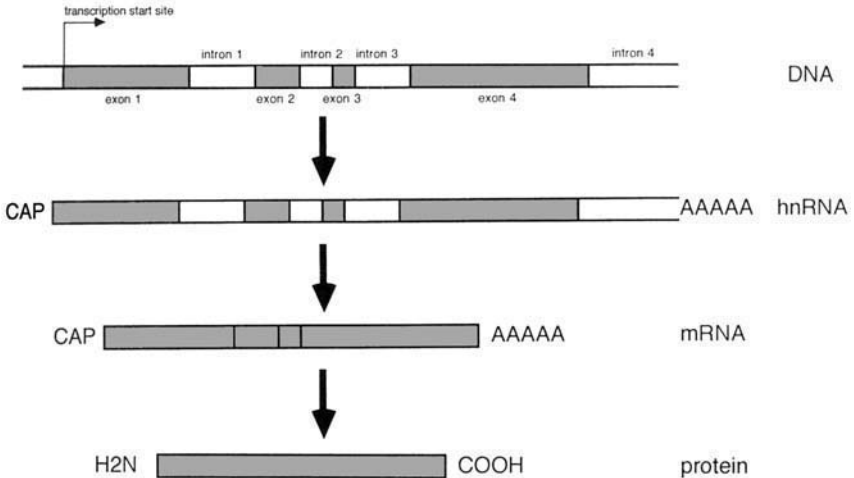


Figure 7. RNA processing. In eukaryotes, many genes are segmented into separate regions called *exons*. Exons are separated by *introns* or regions that do not code for a protein. The RNA polymerase first makes an exact copy of the gene (hnRNA), including the introns. Next, the transcript is processed and leaves the nucleus as mRNA without introns. Finally, during translation, the mRNA is used as the template for the expressed protein (details in text).

tinct from the terminal 5' or 3' untranslated regulatory regions mentioned previously, in that these elements are located within the coding region of the gene. They are removed from the hnRNA molecule, or primary transcript, in the nucleus prior to its transport to the cytoplasm, in a process known as *RNA splicing*, or *RNA processing*. The regions of the DNA that are not spliced (and remain in the mature mRNA molecule) are called *exons* (or expressed regions). Removal of the introns leaves the cell with a final mRNA molecule containing only a series of contiguous exons that will be translated into the final protein. Needless to say, precise intron removal is essential to the production of a translatable mRNA, for the addition or deletion of even a single nucleotide during the splicing process may result in a frame shift mutation and yield a nonfunctional protein.

In the eukaryotic genome, there are examples of both spliced and unspliced genes. Some genes, such as thyroglobulin, are quite large, and possess more than 35 introns. Others, such as mitochondrial genes, are unspliced. Although the function of intervening sequences remains unclear, it is plausible that such regions are remnants of primitive gene segments. Ancient proteins may have evolved by mechanisms such as gene amplification, duplication, or perhaps even by translocation. The introns present in our genome today may provide a link to our primordial heritage.

The process of splicing an immature heterogeneous nuclear RNA

(hnRNA) into a mature cytoplasmic mRNA molecule is under intense investigation. Two mechanisms have been demonstrated and described in some detail: enzyme-mediated gene splicing and RNA-catalyzed splicing (or self-splicing).

Once the primary RNA transcript is made from the DNA template, and the eukaryotic cap has been added to the 5' end along with a 3' poly-A tail, the process of splicing begins. The goal, of course, is to remove extraneous intervening sequences prior to export to the cytoplasm, where translation occurs. The most commonly observed splicing process appears to be enzyme-mediated, using small nuclear ribonucleoproteins (snRNPs). All splice sites (the junctions of introns and exons) have been shown to have special consensus sequences (often called "donor" and "acceptor sites"). These short nucleotide sequences define the "splice sites" of the mRNA, and occur at the intron-exon junctions. In addition, there is a small sequence within the intron that also is used to accomplish the splice. The snRNPs bind to the consensus splice sites (including the internal site), and then associate with one another to form a large complex called the *spliceosome*. It is within this spliceosome structure that the introns are removed and the exons are joined to form a contiguous mature mRNA molecule.^{6, 13}

A newer model for mRNA splicing, described in 1982, earned a Nobel Prize for Thomas Cech^{5a} and Sidney Altman.^{10a} They demonstrated that RNA molecules alone (with no proteins) are capable of catalyzing the splicing reaction. Such a finding created quite a stir among biochemists, who had previously regarded all enzymes as proteins. These RNA molecules, termed *ribozymes*, may act intramolecularly in cis-reactions, where both the catalytic and the substrate domains are on the same RNA molecule, or intermolecularly, where the ribozyme removes introns located on a different RNA molecule. In both cases, the splicing reaction is strongly dependent upon the secondary structure (or folding characteristics) of the RNA target. For example, intramolecular base pairing leads to unusual "stem and loop" configurations that serve to bring one end of an intron in close proximity to the other end, thus allowing for more efficient and precise intron excision.⁵

Thus far, self-splicing reactions have been demonstrated in certain types of viral particles and protozoal species, as well as in yeast mitochondria. Current research on ribozymes is focused on their utility as therapeutic agents in a variety of clinical applications.¹² Many disease processes, such as viral infections (including HIV) and neoplastic disorders, are initiated by the production of an aberrant or mutated gene product. It is hoped that the development of site-specific ribozyme molecules that target certain viral or oncogene RNA sequences may attenuate, or even arrest, the course of the disease. In addition, ribozyme-mediated repair of a defective RNA transcript has been demonstrated in lower organisms. If such a technique can be perfected and applied to higher species, then perhaps the genetic repair of mutations may one day become a reality.

PROTEIN SYNTHESIS

The Genetic Code

The elucidation of the genetic code is a story of perseverance and painstaking laboratory investigation. The genetic code is the Rosetta Stone of molecular biology—once solved, it enabled researchers to predict the amino acid sequence of proteins based upon DNA or RNA sequence data, as well as defining coding regions of genes. Few achievements in biology equal the elucidation of the genetic code in importance.

Once the central dogma had been proposed, workers began to decipher the mechanisms of the translation process. How does genetic information proceed from the mRNA molecule to the full protein? Experiments performed by Brenner and Crick^{8a} in 1961 using bacteriophages (viruses that infect bacteria) showed that each amino acid is described by groups of exactly three adjacent nucleotides. These groups of three nucleotides were called *codons*. The codon hypothesis was reinforced by mathematical permutation analyses that predicted the need for a three-nucleotide system to describe a minimum of 20 possible amino acids. It was left to other researchers, including Marshall Nirenberg and co-workers, using synthetic polynucleotide mRNA strands, an *in vitro* translation system, and protein sequencing techniques, to determine the genetic code.¹⁵ He and his co-workers generated a table of all possible codons and the amino acids that they translated (Table 2).

It is imperative that the order of nucleotides in the DNA template

Table 2. THE GENETIC CODE

First Position (5' End) ↓	Second Position				Third Position (3' End) ↓
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

be preserved. Mutations consisting of replacements, deletions, or additions will not only affect the codon in which they occur, but also may shift the “reading frame” of all codons downstream of the mutation. For example, the RNA sequence . . . AGG-GCC-UAC . . . will ultimately produce a protein with the amino acid sequence: arginine-alanine-tyrosine. The addition of one base at position 2 will result in an RNA sequence of: . . . AAG-GGC-CUA-C . . . , and a new amino acid sequence of lysine-glycine-leucine. Similarly, it is easy to read the phrase “the boy ran. . . ,” but if we delete the third letter (yet maintain a three-letter “codon”), it becomes: thb oyr an. . . , a tough sentence to read. Hence it is important to maintain both an accurate sequence as well as a conserved reading frame.

Translation

Following transcription and RNA splicing in the nucleus, the mature mRNA molecule conveys selected genetic information to the cytoplasm where protein synthesis (translation) occurs. It is this message that will instruct the *ribosomes*, complex macromolecules composed of rRNA and proteins, exactly how to join amino acid residues together in order to form the protein molecule. Thus, the four-letter alphabet of the nucleotides (A, U, C, or G) is translated into the 20-amino acid alphabet of the protein. Every three consecutive bases in the mRNA are considered as one unit (the “codon”). Each codon codes for a specific amino acid (or a stop or start signal). Once in the cytoplasm (the worksite in the construction analogy), the mRNA convenes the translational cast of characters, including ribosomes, transfer RNA (tRNA) molecules with their attached amino acids, initiation factors, elongation factors, termination factors, and energy-producing molecules (GTP and ATP), for what is to be their Final Act. This finale occurs in three scenes: initiation, elongation, and termination (with post-translational modification of the protein product constituting the epilogue in certain instances).

The translation of the mRNA into a protein requires a translator; that is, a molecule fluent in both languages. This function has been assumed by the tRNA molecule. It serves as an adaptor molecule, able to recognize codons on the mRNA sequence as well as the corresponding amino acid that it “codes for.” Every tRNA molecule is pre-ordered (by nature of its anticodon loop) to carry only one amino acid: the correct amino acid for its particular anticodon, as dictated by the genetic code. An enzyme called aminoacyl-tRNA synthetase is responsible for putting the right amino acid on each tRNA molecule. There are 20 types of tRNA synthetase enzymes, one for each amino acid, just as there are 20 types of tRNA molecules. It is the tRNA molecule that recognizes the codon on the mRNA; the amino acid plays no role in this recognition process.

During initiation, the mRNA latches onto the smaller of the two ribosomal units and displays its initiation codon (AUG). A tRNA des-

tioned to couple with the initiation codon (by virtue of a UAC anticodon that recognizes the complimentary AUG sequence) does so, thus ensuring that its amino acid, methionine, will be the first amino acid in the peptide chain. At this point, the larger ribosomal unit recognizes the mRNA–smaller ribosome–tRNA with methionine complex and, with the energy resulting from the hydrolysis of GTP, joins the initiation complex. This complex also has a supporting cast of perhaps a dozen or so different initiation factors, each assisting in some aspect of the assembly of the initiation complex.

With the two (large and small) ribosomes joined in the initiation complex, elongation can proceed (Fig. 8). Each nucleic acid triplet subsequent to the initiation codon prompts the recognition by an anticodon belonging to a tRNA whose other end contains a specific amino acid. With the nucleic acid portion of the tRNA thus bound to the mRNA, the amino acid from the new tRNA lines up with the previous tRNA's amino acid, and these two amino acids are promptly fused by the action of the enzyme peptidyl transferase. The ribosome complex then migrates down the mRNA to the next codon, releasing the older, original tRNA (now devoid of the growing polypeptide chain), and inviting a new tRNA partner to join the party—a move termed *translocation*. In this manner, amino acid after amino acid is added to the growing peptide chain, as directed by the sequence of triplet bases in the mRNA.

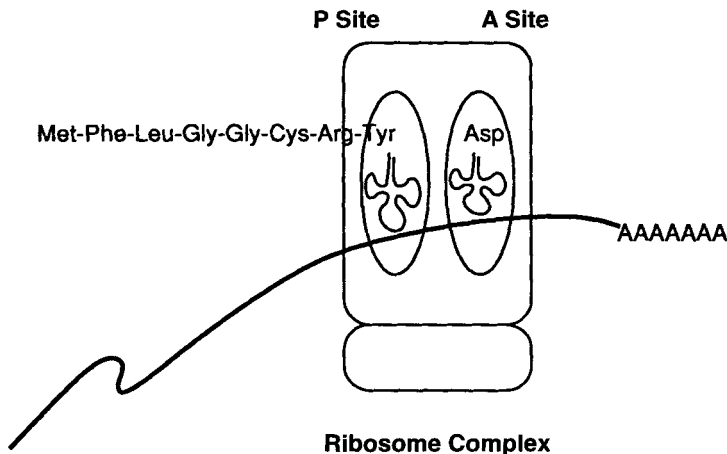


Figure 8. The translation process. The process of translation (as discussed in the text) is basically the conversion of a chain of mRNA codons into a chain of peptides (amino acids). Translation uses the tRNA molecule as the “go-between,” recognizing the mRNA codon with its anticodon sequence and bringing along the correct amino acid on its back. The tRNA bearing the next amino acid directed by the mRNA enters the ribosome at the A-site and accepts the growing polypeptide chain in a transfer from the tRNA in the P-site (using peptidyl transferase). This leaves the empty tRNA in the P-site with nothing to do, so it vacates the site. The ribosome then moves over one codon (translocation), transferring the tRNA bearing the polypeptide to the P-site, and the process begins again.

The ribosomes actually contain two tRNA binding sites, P and A. The incoming tRNA-amino acid (the one to be added to the peptide chain) enters at the A site. After the tRNA has attached to the mRNA (by codon-anticodon hydrogen bonding), and the new amino acid has joined the polypeptide chain (using peptidyl transferase), the ribosome translocates, moving the newest amino acid to join into the P site, while leaving the A site open for the next tRNA-amino acid. The A and P notations stand for amino-acyl and peptidyl, respectively.

After the last amino acid has been recruited, the mRNA codes for the termination of the peptide chain. This is signaled by one of the three stop codons in the genetic code (UAA, UAG, or UGA). There are no amino acids associated with these codons. These termination codons result in the release of the newly synthesized protein and dissociation of the ribosomal subunits and mRNA.

In nature's efficient design, a single mRNA can direct the synthesis of several chains of protein concurrently, with multiple ribosome-tRNA complexes bound to a single mRNA at various points along the message strand. Also, it should be noted that once the amino acids have been joined in a polypeptide chain, and that chain has assumed its natural three-dimensional configuration (native conformation), even further modification of the protein can occur. Examples of such modifications include glycosylation, acetylation, methylation, or disulfide bridge formation. In addition, the proteins may be cleaved into an active form, or combined with other proteins to form large complexes.

Regulation of Gene Expression

Perhaps the most important aspect of gene expression is the intricate system of regulation that has evolved through countless generations of life on earth. Nature has devised an amazing interrelated system for the storage, transcription, translation, and duplication of genetic material, yet these mechanisms pale in comparison to the elegant control systems used to regulate these processes. There are several types of regulation commonly used by eukaryotic organisms. The choice of regulatory system is specific for each particular gene, and often depends upon the needs of the cell. For example, there are many gene products (enzymes or structural proteins) that are needed continuously by the cell. These are often referred to as "housekeeping genes," reflecting their essential role in cellular structure or metabolism. Most housekeeping gene products are expressed constitutively, meaning that they are always "turned on," or transcriptionally active. Other genes are inducible, meaning that they are activated only at specific times in development or in response to specific stimuli. It would be wasteful to transcribe and translate these gene products when they are not needed. Yet gene regulation has

evolved beyond the simple on-off switch for gene expression that is used by prokaryotes. In eukaryotes, many levels of control are used for inducible genes, including transcriptional, post-transcriptional, and translational regulation. We briefly mention primitive regulatory systems, and then explore these advanced schemes.

Most early studies of gene regulation focused on prokaryotic systems. Nearly all gene regulation in prokaryotes occurs at the transcriptional level. Elegant work by Jacques Monod and Francois Jacob^{10b} led to the discovery of the operon motif in *Escherichia coli*. This grouping of several related structural genes is activated by inducing agents that interact with and remove a repressor molecule from the DNA upstream control region (the promoter) of the operon. Once removed, RNA polymerase can bind to the promoter at the site previously blocked by the repressor, and transcription can occur. In the case of the lactose operon in bacteria, the lac operon is maintained in an "off" state (transcriptionally silent) until there is ample substrate (lactose) available, and thus a physiologic need for the operon's gene products (lactose-metabolizing enzymes). Positive regulation also can occur in prokaryotes, as when an activator protein binds to the DNA and stimulates transcription of certain genes. The operon motif can be activated and deactivated rapidly in response to changes in the environment, thus ensuring that transcription of metabolic enzymes occurs only when adequate amounts of substrates are available. Although eukaryotic gene regulation is far more complex, the economic principle of gene transcription only when needed (as in prokaryotes) still holds true.

There are several mechanisms through which the regulation of eukaryotic gene expression occurs. Ultimately, it is the production of a complete, functional protein product that is the true measure of gene expression. Genes may be regulated at the *transcriptional* level, where differential rates of the transcription of a gene in certain cell types or at certain times of development occur. An accumulation of a particular mRNA can lead to its increased translation and, therefore, an increased production of the protein it describes. The chromatin structure of the DNA molecule also influences the transcription of certain genes. Prior to transcription, the DNA must unwind, and the nucleosomes must allow transcription factors and RNA polymerase to access their recognition sequences. *Post-transcriptional* regulation is manifested primarily by alterations in mRNA stability. There are certain structural motifs that make an mRNA molecule more susceptible to RNase degradation, and thus relatively short-lived. In this instance, the amount of translation that can occur is limited by the half-life of the mRNA. Alternative splicing of the primary transcript also may regulate the expression of a gene product by producing different mature mRNA molecules. Many eukaryotic genes use alternative splicing to produce multiple proteins from a single gene. The portion of the immunoglobulin molecule that determines whether the molecule will be secreted or bound to the membrane is regulated in this fashion. *Translational* regulation is an

infrequently used mechanism of gene regulation, yet in certain cases selective modification of translation initiation factors can slow or halt the translation of certain mRNAs. Negative translational control also is present in eukaryotes, wherein the translation of some mRNA molecules is blocked by translation repressor proteins. *Post-translational* regulation also can occur. This type of regulation commonly is observed among proteins that require some level of processing or cleavage before assuming an active, functional conformation.

In eukaryotes, most of the regulation of gene expression occurs at the transcriptional level. The decision of what and when to transcribe is perhaps the most important decision a cell can make (figuratively speaking, of course). The area of transcriptional regulation is so exciting and popular at this time that it merits a more detailed discussion. We discuss the promoter region as well as other DNA regulatory elements.

There are basically two components to each regulatory region on a DNA template: a recognition sequence and a protein that binds to it. In the case of transcription, the presence of both elements usually results in efficient transcription. We previously have introduced the promoter region, with its several consensus sequences located immediately upstream of the region to be transcribed. Exhaustive sequencing of a variety of eukaryotic genes revealed two "classes" of promoter sequences (Fig. 9A). The first type were those sequences found in nearly all the genes, and probably involved in general transcription of genes that are expressed routinely in all cells. These "constitutive" promoter elements, such as the nearly ubiquitous TATA box, function as adequate, albeit weak, promoters when present alone. The addition of other elements (sequences) upstream to these weak promoter sequences, however, turn them into strong promoters. These other upstream sequences, typically within 200 bases of the first transcribed nucleotide, are found in large numbers of genes (although not as many as the consensus TATA sequence). They have been given abbreviated names, such as Sp1 or AP-1, often reflecting where they were first identified. A typical gene may possess six or more of these sequences (excluding the TATA or CAAT boxes), or even multiple copies of the same element.

The second type of promoter element identified was found only in genes transcribed in a particular tissue, or in response to a specific stimuli. These "inducible" promoters seem to have sequences that are shared only by a small number of identified genes. They are often found mixed among the constitutive elements in the upstream region of a gene. For example, all genes containing the AP-2 sequence are activated in response to cyclic AMP or phorbol esters. This inducible type of promoter is also responsible for tissue-specific transcription. Often, the protein factor that binds to these inducible elements is restricted to a special cell type. For example, within the promoter region for the immunoglobulin genes (heavy and light chain) there is a sequence of eight nucleotides called the "octomer" motif. The sequence works by binding a protein called Oct-2, which is absent in most cells except for immunoglobulin-producing B cells. These specialized proteins that bind

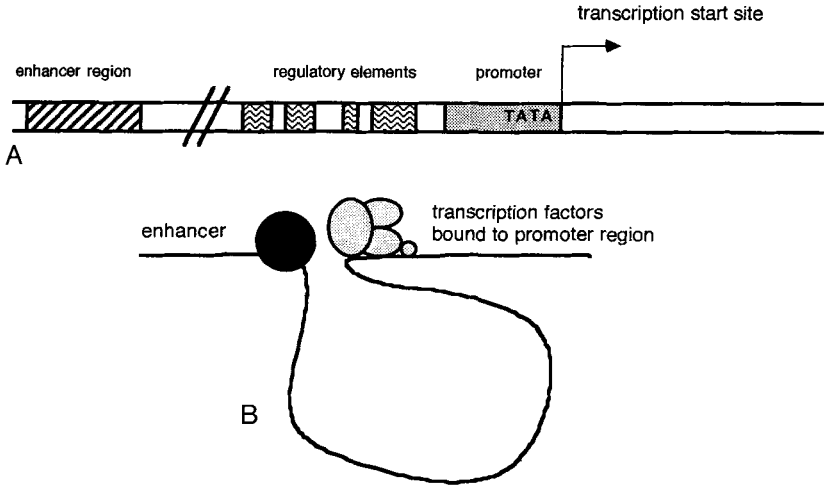


Figure 9. Transcriptional control regions. The 5' or upstream region of the gene contains *noncoding sequences responsible for the regulation of transcription*. The promoter region, usually within 30 bases of the transcription start site, contains fairly common regulatory sequences that help identify the start site. Further upstream, gene-specific regulatory elements (sequences) are found. These often contain binding sites for transcription factors that are specific for a family of genes or certain tissue type. The enhancer region is located at a distant site from the gene and presumably activates gene transcription by binding protein factors and then associating with the promoter region. This is facilitated by large loops formed in the DNA molecule.

to promoter elements and activate the gene transcription are called *transcription factors*.

Another type of regulatory element that must be mentioned is the *enhancer*. An enhancer consists of interesting, complex DNA sequences that act to increase the activity of a specific promoter in a location hundreds, or even thousands of nucleotides away. In this respect, they may be regarded as a transcriptional "remote control" system. Furthermore, the enhancer is capable of functioning perfectly well in locations that are upstream, downstream, or even within the gene they are controlling. In addition, they are active in either orientation relative to the gene. How does the enhancer regulate genes at such a great distance, and in either orientation? The answer lies in the ability of DNA to bend and assume a complex secondary structure. Like promoters, enhancers contain multiple binding sites for transcription factors. The DNA between the enhancer and the promoter of a gene forms a large loop, allowing the enhancer's transcription factors to interact with either the promoter's transcription factors or with the RNA polymerase itself. The DNA acts like a tether, keeping the enhancer and its bound transcription factor in close association with the promoter region of the gene (Fig. 9B).

MOLECULAR BIOLOGY TOOLS

Restriction Enzymes

As previously mentioned, eukaryotic DNA exists in a tightly condensed structure of nucleic acid wound around histone and other nucleoproteins called chromatin. It is quite difficult to analyze individual genes with the DNA in this form. In the 1970s, W. Arber,^{1a} D. Nathans, H. Smith,^{16a} and co-workers isolated enzymes called *restriction endonucleases* from several bacterial strains. In bacteria, these restriction enzymes form a primitive immune system—defending the bacteria from invading bacteriophages (viruses that infect bacteria) by cleaving, or digesting, foreign DNA. Restriction enzymes digest double-stranded DNA molecules at specific sequences, thus introducing strand breaks into the invading DNA. The recognition sequences are usually short (3–8 nucleotides) and occur randomly throughout the genome at frequencies that may be estimated mathematically based upon the length of the recognition sequence. The bacteria protects its own DNA from digestion by modifying the recognition sequences so they are not digested (perhaps the first case of self-non-self discrimination).

Molecular biology has taken advantage of these enzymes and used them as tools for genetic analysis. Purified restriction enzymes are available commercially, and are used like molecular scissors to cleave DNA molecules at specific sequences. The cleaved DNA is easier to analyze, or fractionate, and resolve using electrophoresis. Because restriction enzyme sites occur at random locations in the genome, there is no discrimination between coding and noncoding regions of the DNA. Thus, the structural (coding) region of a gene may be fragmented with a particular restriction enzyme. Fortunately, there are hundreds of restriction enzymes available with different defined recognition sequences, so an appropriate one may be found for every application.

Gel Electrophoresis

The technique of electrophoresis is not unique to the study of nucleic acids. For decades, this method has been used to fractionate proteins and other macromolecules. In the past, matrices such as cellulose acetate strips were used in the analysis of proteins. More recently, the use of polyacrylamide polymers and agarose of varying concentrations have proven invaluable to the analysis of nucleic acids. We briefly discuss the principles of gel electrophoresis, and then explore the many and varied uses of this technology.

Electrophoresis is a technique used to separate mixtures of macromolecules based both on molecular weight and electrical charge. The mixture of soluble molecules is placed in a trough, or well, located at the top, or origin, of the gel. The sample is commonly “loaded” onto the gel along with a marker dye solution, which enables the researcher

to follow the progress of the separation. An electrical current then is applied to the gel, with the positive terminal (the anode) at the bottom of the gel. This arrangement takes advantage of the net negative charge conferred upon nucleic acids because of the negative charge of each nucleotide. The speed at which the molecules migrate through the gel is also dependent upon the overall size of the molecule—that is, the larger the molecule, the slower its migration through the gel matrix. Typically, multiple samples (including a set of defined size markers) are electrophoresed (or “run”) on one gel at a given voltage for times ranging from several hours to overnight. These conditions have been optimized to maximize resolution (the ability to discern several species of similar molecular weights).

Following electrophoresis, various techniques are used to visualize the nucleic acids. Intercalating dyes, such as ethidium bromide, commonly are applied. The completed gel is soaked in a solution of this compound, then visualized under ultraviolet lighting. The dye molecules lodge in between the base pairs of the nucleic acid and fluoresce following ultraviolet excitation. A photograph is taken and can be used to assess nucleic acid size, relative concentration, integrity, and purity. In the case of preparative analysis of nucleic acids, the procedure stops here, but for further detailed analytical studies, and to definitively identify the observed bands on the gel, a further step is undertaken. This next step involves transferring the electrophoresed molecules to a solid matrix, such as a nitrocellulose or nylon filter, and is known as “blotting” the gel.

Blotting and Hybridization

The general technique of gel transfer to a solid matrix by capillary action was developed by E. Southern in 1975, and to this day bears his name.¹⁷ The Southern blot is typically performed with DNA molecules. A similar procedure used for analysis of RNA molecules is known as “Northern” blotting, no doubt a reflection of the irrepressible sense of humor common to all molecular biologists. The “Western” blot involves the transfer of electrophoresed proteins, usually through a polyacrylamide gel, to a membrane or filter.

Following electrophoresis and staining, the gel typically is soaked in a denaturing solution to “open up” the DNA helix and to make the bases accessible for later hybridization reactions (RNA gels do not require this step). The transfer filter (either nitrocellulose paper or a nylon membrane) is placed atop the gel, and the transfer is accomplished by capillary transfer overnight using a dish filled with hypertonic salt solution. Typically, the capillary action is achieved by stacking paper towels or tissues on top of the membrane filter. This allows the salt solution to be sucked up first through the gel, then the membrane, and finally into the towels. The filter is then baked for a short time to

permanently "fix" the nucleic acid to the membrane, and is then ready for hybridization.

Newer technology has been developed to speed up the transfer procedure using either a vacuum apparatus in place of the capillary action, or even an electrical current to rapidly transfer molecules from the gel to the solid matrix. The speed of transfer can then be reduced to 2 hours or so.

Once the blot has been performed, the filters are quite stable and may be stored for later use. Owing to the competitive nature of the present funding situation, however, most blots are used immediately. In the case of nucleic acid blots (Northern and Southern), the membrane typically is annealed (or "hybridized") to a radiolabeled "probe." The probe is of a defined sequence, usually isolated as either a cloned or enzymatically synthesized fragment using recombinant DNA techniques. Once isolated, it is radiolabeled in a polymerase reaction using one or more ^{32}P -tagged nucleotides. The probe then is purified to remove unincorporated label, and added to a bag or bottle containing the blot and a suitable solution that favors the annealing of the probe to a homologous sequence somewhere on the blot. After a short time (from 2–24 hours), the blot is washed under defined conditions of temperature and tonicity to remove nonspecifically bound probe. These conditions also can be adjusted to maintain high stringency, thus ensuring that the probe will only bind to sequences of highest homology. The final step is to expose the hybridized and washed blot to radiographic film. Regions that are homologous to the probe now appear as dark bands. They can be sized accurately and further analyzed. Recently, in an effort to reduce exposure to high levels of radioactivity, several companies have developed nonradioactive probe labeling systems based on either enzymatic or fluorescent reactions.

The Polymerase Chain Reaction

The polymerase chain reaction (PCR) is a recently described method used to amplify low-abundance DNA sequences. It would be an understatement to say that this technique has revolutionized the field of molecular biology. Its use in the research laboratory is now commonplace, but perhaps more surprising is the breadth of varied disciplines that have embraced this technology. Forensic medicine uses PCR extensively to amplify DNA sequences from a few drops of blood at a crime scene, for example. Infectious disease laboratories use the technique to identify certain pathogens in tissue or blood samples. The identification of *Mycobacterium tuberculosis* organisms in sputum samples, for example, could save the 6 weeks of culture time needed to verify a tuberculosis diagnosis. Medical genetics labs use PCR methodology to investigate inheritable diseases during the prenatal period, using minute amounts of tissue (often several cells' worth of DNA is sufficient). If the 1980s

were the era of the Southern blot and gene cloning, then surely the 1990s are the era of the PCR reaction.

The PCR reaction takes advantage of several features. As noted previously, DNA polymerase is capable of copying a single-stranded template to make a complementary new strand. The reaction requires a double-stranded starting point, or initiator region, upon which the polymerase can add nucleotides (similar to having to "start" a zipper by hand prior to the action of "zipping"). The starting point in this case is called a "primer," or short piece of DNA (10–30 nucleotides) that binds to the template and tells the polymerase where to start copying.

The PCR technique, originally devised by Kary Mullis,¹⁶ takes advantage of a unique, thermostable DNA polymerase (*Taq* polymerase) isolated from the bacterium *Thermus aquaticus*. This bacterium is unusual because it thrives in temperatures exceeding 75°C (such as those found in natural geysers). The DNA polymerase from this strain performs optimally at 72°C, but is even stable at temperatures exceeding 90°C, where most double-stranded DNA molecules are effectively denatured into single strands. Thus, the *Taq* polymerase added at the beginning of a reaction remains active through a wide range of temperatures.

The PCR reaction uses oligonucleotide primers, usually synthetic DNA sequences of approximately 20 bases, which flank the region of interest. Therefore, it is imperative that the sequence of interest be known, or at least inferred from the target protein amino acid sequence. Thus, if we sought to amplify a sequence from *M. tuberculosis* to determine its presence in a sputum sample, we might first make (or purchase) a pair of primers that are complementary to a portion of the genome. Ideally, a primer should be unique enough to decrease possible cross-reactivity. The second primer forms the second boundary of the flanking region. It is between these two selected primers that the *Taq* polymerase will act to copy the template.

The actual PCR reaction is done in a tube containing the DNA "template" (containing the sequence to be amplified), the *Taq* polymerase, a large supply of the two primers (one for each side of the sequence of interest), free nucleotides to make the amplified DNA, and other factors used for polymerization. The mixture initially is heated to a high temperature to denature the template DNA into single strands, then taken to a lower temperature to allow the primers to anneal to the template (there is such a large excess of primers, that the parent strands cannot find each other to reanneal) and replication to occur. After a short time, the temperature is raised again to separate the newly formed DNA (with the primers attached), and the process is repeated (often up to 30 times). The number of synthesized molecules can be approximated because of the geometric progression of the amplification. By the time 30 cycles have been completed, more than one billion copies of the sequence of interest will have been made. This is usually enough to satisfy even the most gluttonous of molecular biologists.

Reverse PCR (RT-PCR)

Reverse PCR (or reverse-transcribed PCR) is a powerful variation of PCR technology. Whereas PCR uses specific primers to amplify DNA sequences directly, RT-PCR amplifies transcribed RNA sequences by first reverse-transcribing the RNA molecules into DNA sequences (using the reverse-transcriptase enzyme mentioned previously), and then amplifying the DNA using the PCR reaction as described. The additional step of reverse transcription allows for the detection and even quantitation of mRNA molecules in a cell or tissue. The usefulness of such a technique is apparent in the examination of cytokines produced by cells obtained by bronchoalveolar lavages (BALs) from asthmatic individuals. Often, even such sensitive techniques as radioimmunoassays are unable to quantitate the minute amount of cytokine protein produced by the small number of cells obtained. As there is presently no refined methodology that allows for amplification of protein molecules to increase sensitivity, amplifying the mRNA coding for these cytokines is the next best alternative. Obviously, the usefulness of the technique is limited to those genes that are regulated transcriptionally (thus far, the majority of known genes).

As with all PCR techniques, the use of appropriate controls and standards is essential to the accurate interpretation of the results. In RT-PCR reactions, special controls are needed to detect the presence of any genomic DNA that may be contaminating the RNA preparation. Extraneous DNA would be amplified during the PCR step of the RT-PCR reaction, and produce a falsely elevated estimation of mRNA concentrations. Recently, a number of approaches to the quantitation of the mRNA, most based upon amplification of known quantities of standards, have been described.

CONCLUSION

Nearly all fields of basic science and medicine have benefited from the exponential growth of molecular biology over the past few decades. The elucidation of the molecular mechanisms responsible for a disease state will enhance its early detection, and possibly lead to novel and effective therapies. The modern clinician must seek to understand both the principles and uses of this powerful technology.

Historically, we sought to understand allergic diseases such as asthma using basic (yet essential) tools including the history and the physical examination. Further investigations led us to objective measures of pulmonary function testing, and even invasive tests such as bronchoscopy. We have examined cell populations in bronchoalveolar lavage samples, and even sampled the airways of asthmatic patients. But only now are we daring to explore the genetic mechanisms that contribute to airway inflammation. Sophisticated studies that examine the expression of cytokine genes in airway cells, or the molecular mechanisms of

steroid resistance, are now the norm. The future promise of allergy and immunology lies among the innermost workings of the cell. Molecular biology is the vehicle that can take us there.

References

1. Alberts B, Bray D, et al: *Molecular Biology of the Cell*. New York, Garland Publishing, 1994
- 1a. Arber W, Linn S: DNA modification and restriction. *Annu Rev Biochem* 38:467–500, 1969
2. Avery OT, MacLeod CM, et al: Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J Exp Med* 79:137–158, 1944
3. Baltimore D: Viral RNA-dependent DNA polymerase. *Nature* 226:1209–1211, 1970
4. Berk AJ, Sharp PA: Structure of the adenovirus 2 early mRNAs. *Cell* 14:695–711, 1978
5. Cech T, Uhlenbeck O: Hammerhead nailed down. *Nature* 372:39–40, 1994
- 5a. Cech TR, Zaug AJ, Grabowski PJ: In vitro splicing of the ribosomal RNA precursor of tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* 27:487–496, 1981
6. Chabot B, Black DL, et al: The 3' splice site of pre-messenger RNA is recognized by a small nuclear ribonucleoprotein. *Science* 230:1344–1349, 1985
7. Chargaff E: Chemical specificity of nucleic acids and mechanisms of their enzymic degradation. *Experientia* 6:201–209, 1950
8. Chow LT, Gelinas RE, et al: An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12:1–8, 1977
- 8a. Crick FH, Brenner S: General nature of the genetic code for proteins. *Nature* 192:1227–1232, 1961
9. Darnell JE, Lodish H, et al: *Molecular Cell Biology*. New York, Scientific American Books, W.H. Freeman, 1990
10. Griffith F: The significance of pneumococcal types. *Journal of Hygiene* 27:141–144, 1928
- 10a. Guerrier-Takada C, Gardiner K, Marsh T, et al: The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* 35:849–857, 1983
- 10b. Jacob F, Monod J: Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318–356, 1961
11. Judson HF: *The Eighth Day of Creation*. New York, Simon & Schuster, 1979
12. Kiehnopf M, Esquivel EL, et al: Clinical applications of ribozymes. *Lancet* 345:1027–1031, 1995
13. Lerner MR, Steitz JA: Are snRNPs involved in splicing? *Nature* 283:220–224, 1980
14. Lewin B: *Genes V*. New York, Oxford University Press, 1994
15. Matthaei JH, Nirenberg MW: Characteristics and stabilization of DNase-sensitive protein synthesis in *E. coli* extracts. *Proc Natl Acad Sci USA* 47:1580–1588, 1961
16. Saiki RK, Gelfand DH, et al: Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239:487–491, 1988
- 16a. Nathans D, Smith HO: Restriction endonucleases in the analysis and restructuring of DNA molecules. *Annu Rev Biochem* 44:273–293, 1975
17. Southern EM: Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503–517, 1975
18. Watson JD, et al: *Recombinant DNA*. New York, Scientific American Books, W.H. Freeman, 1992
19. Watson JD, Crick FHC: Genetic implications of the structure of deoxyribonucleic acid. *Nature* 171:964–967, 1953
20. Watson JD, Crick FHC: Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* 171:737–738, 1953

Address reprint requests to

Randy J. Horwitz, MD, PhD
 Section of Allergy and Immunology
 University of Wisconsin Hospital & Clinics
 600 Highland Avenue, Room H6/367
 Madison, WI 53792