APPENDIX

THE CHEMISTRY OF LIFE

This appendix will give the interested reader an overview of biochemical principles that undergird life. It is not necessary to read the appendix to follow the arguments in the book, but it will set those arguments within a larger framework. Here I will discuss cells and the structures of several major classes of biomolecules—proteins and nucleic acids and, briefly, lipids and carbohydrates. I will then focus on the question of how genetic information is expressed and propagated. Of course, in such a short space the description must be sketchy, so I urge those who become intrigued by the mechanisms of life to borrow an introductory biochemistry text from the library. A fascinating Lilliputian world awaits.

CELLS AND MEMBRANES

The human body is composed of hundreds of trillions of cells. Other large animals and plants also are conglomerations of enormous numbers of cells. As the size of an organism decreases, however, the number of cells decreases also; for example, the small worm *C. elegans* contains only about a thousand cells. As we travel down the size scale we ultimately reach the unicellular phyla, such as yeast and bacteria. No independent life occurs below this level.

Examination of its structure shows why the cell is the fundamental unit of life. The defining feature of a cell is a membrane—a chemical structure that divides the outside world from the interior of the cell. With the protection of a membrane, a cell can maintain different conditions inside than prevails outside. For example, cells can concentrate nutrients in their interior so that they are available for energy production, and can prevent newly made structural materials from being washed away. In the absence of a membrane, the large array of metabolic reactions necessary to sustain life would quickly dissipate.

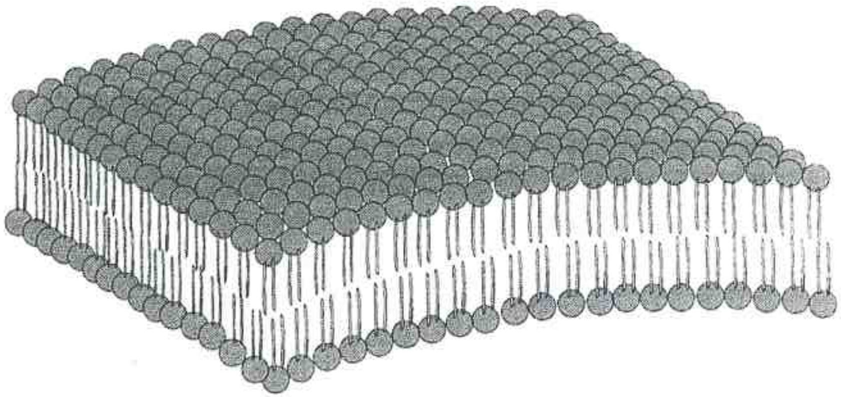
Cell membranes are made from amphiphilic molecules that are similar in ways to the soaps and detergents used in household cleaning. The word *amphiphilic* is from the Greek meaning "loves both"; an amphiphilic molecule "loves" two different environments: oil and water. The shape of the molecules is roughly similar to a lollipop with two sticks coming out the same side of the candy ball. The sticks usually consist of hydrocarbons (made from atoms of carbon and hydrogen) and, like other hydrocarbons such as gasoline, do not mix well with water. This is the oil-loving part of the molecule. Such regions of molecules are called *hydrophobic,* from the Greek for "water-fearing." The ball of the lollipop molecule, in contrast, generally has a chemical group that, like table salt or sugar, positively enjoys being in water. Such regions are called *hydrophilic* ("water-loving"). The two opposite parts of membrane molecules are chemically tied together and, like Siamese twins, must travel together despite dissimilar properties. But if one part of the molecule wants to be in water and the other part wants to be out of water, where does the molecule settle down?

Amphiphilic molecules solve their dilemma by associating with other amphiphilic molecules. When a large number of amphiphiles associate, the hydrophobic tails all huddle together to exclude water while the hydrophilic heads touch the water. An efficient way for the tails to be shielded from water while still allowing the water-loving groups access to water is to form two sheets (Figure A–1), called a *lipid bilayer.* If the two sheets remained flat, however, the hydrocarbons at the edges of the sheets would remain exposed to water. So the sheets close up, like a soap bubble.

Since the middle of the membrane bilayer is oily, many molecules that strongly prefer a watery environment (such as salts and sugars) cannot cross the membrane. Thus we have a structure with an enclosed interior that can be different from the outside environment—the first step in making a cell.

FIGURE A–1

A SEGMENT OF A LIPID BILAYER.



The living world contains two fundamentally different type of cells: the *eukaryotes,* in which a second membrane, different from the cell membrane, encloses the nucleus of the cell; and the *prokaryotes,* which do not have this feature.[1] Prokaryotic organisms are invariably unicellular and are, in many ways, much simpler than eukaryotes.

Besides the cell membrane only a few features stand out in photographs of prokaryotes.[2] One is the *nucleoid,* the mass of cellular DNA (deoxyribonucleic acid) resting comfortably in the middle of the *cytoplasm* (the soluble cell contents). In addition to a membrane, prokaryotes have a second structure surrounding the cell, called the *cell wall.* Unlike the membrane, the cell wall is made of polysaccharide that is rigid and freely permeable to nutrients and small molecules. It confers mechanical strength, preventing the cell from rupturing under stress. Several structures stick out from the membrane of many prokaryotic cells. The function of the hairlike *pili* is largely unknown. The bacterial *flagellum* is used for locomotion; flagella rotate rapidly like a propeller to move the prokaryote along.

The second category of cells is the eukaryotes, which compose all multicellular organisms, as well as some single-celled organisms like yeast. Eukaryotic cells contain a number of subcellular spaces that are separated from the cytoplasm by their own membranes; these are called *organelles*, because they are reminiscent of the organs found in the body of an animal. Organelles allow the eukaryotic cell to conduct specialized functions in specialized compartments.

The first specialized organelle is the *nucleus*, which contains the cell's DNA. The membrane surrounding the nucleus is a highly specialized structure, perforated by large, eight-sided holes called *nuclear pores*. The pores are not passive punctures, however; they are active gatekeepers. No large molecule (like proteins or RNA) gets past the nuclear pores without the correct "password." This keeps molecules that belong in the cytoplasm out of the nucleus, and vice versa.

A number of other organelles stud the cytoplasm. *Mitochondria* are the "power plants" of the cell; they specialize in the chemical reactions that turn calorie-laden nutrient molecules into forms of chemical energy that the cell can use directly. Mitochondria have two membranes. The controlled "burning" of nutrient molecules generates a difference between the acidity of the space enclosed by the inner membrane and that enclosed between the inner and outer membranes. The controlled flow of acid between the two compartments generates energy, like the flow of water over a dam generates electrical power.

Lysosomes are small organelles bounded by a single membrane; essentially, they are bags of enzymes which degrade molecules that have outlived their usefulness. Molecules destined to be degraded in the lysosomes are transported there in small, coated vesicles (see Chapter 5). The acidity in the lysosome is one hundred to one thousand times greater than that in the cytoplasm. The increased acidity makes tightly folded proteins open up, and the open structures are then easily attacked by degradative enzymes.

The *endoplasmic reticulum* (ER) is an extensive, flattened, convoluted membrane system that is divided into two different components: the rough ER and the smooth ER. The rough ER gets its craggy appearance from numerous *ribosomes* attached to it; ribosomes are the cellular machinery that synthesize proteins. The smooth ER synthesizes lipids—fatty molecules. The *Golgi apparatus* (named for Camillo Golgi, who first observed it) is a stack of flattened membranes to which many proteins made in the ER go for modification.

A cell can take on shapes radically different from spherical (for example, a sperm cell), and can change shape in response to changes in the environment. The shape of the cell is supported by the *cytoskeleton*, which, as its name implies, is the cell's structural framework. The cytoskeleton is composed of three major structural materials: microtubules, microfilaments, and intermediate filaments. *Microtubules* serve a number of functions. Among these are formation of the mitotic spindle—the apparatus that, during cell division, pushes one copy of each chromosome into each daughter cell. Microtubules are also the spine of eukaryotic cilia, which, like oars, can move the cell through its environment. Finally, microtubules can act as "railroad tracks" for molecular motors to carry cargo to distant parts of the cell. *Microfilaments*, thinner than microtubules, are made of the protein actin, which is also a major component of muscle. Microfilaments grab onto each other and slide to contract. This shapes the cell by folding the cellular membrane at the right places. *Intermediate filaments*, which are thicker than microfilaments but thinner than microtubules, seemingly act simply as structural supports (like steel girders). Intermediate filaments are the most diverse structures of the cytoskeleton.

Almost all eukaryotic cells contain the organelles described above. Plant cells, however, contain several additional organelles. The *chloroplast* is the site of photosynthesis. Chloroplasts are, in many ways, similar to mitochondria since they both have energy-generating responsibilities. Chloroplasts contain the pigment chlorophyll, which acts as an antenna to catch light. The energy of the light is passed to extremely complex molecular machinery that generates differences in acidity across the membranes of the chloroplast. Plant cells also have a large, clear, membrane-enclosed space called the *vacuole*. The vacuole is a reservoir for wastes, nutrients, and pigments, and it also has a structural role. The vacuole occupies about 90 percent of the volume of some plant cells and is under high osmotic pressure. The pressure, pushing against a strong plant cell wall, stiffens the cell.

PROTEIN STRUCTURE

The cells and organelles described above, although quite tiny by everyday standards, are very large compared to the building materials of which they are composed. The building materials of cells and subcellular structures are ultimately composed of *atoms* stitched together into *molecules*. A chemical bond, or *covalent* bond, forms when each of two atoms contributes an electron to share between them. By sharing negatively charged electrons, the atoms more efficiently screen their positively charged atomic nuclei. A molecule is two or more atoms covalently bonded to each other.

Surprisingly, the types of atoms found in biological molecules are few. Almost all biomolecules are made of atoms of six elements: carbon (C), oxygen (O), nitrogen (N), hydrogen (H), phosphorous (P), and sulfur (S). Some other elements (such as chlorine, sodium, calcium, potassium, magnesium, and iron) are found as ions in biological systems. (Ions are electrically charged particles that float more or less freely in water.)

Atoms of C, H, O, N, P, and S can bond with each other. Carbon can bond with up to four different atoms at once, and biological phosphorus can also bond four different atoms (almost always four oxygens). Nitrogen can form three bonds (four in special cases), and oxygen and sulfur can form two. Hydrogen can form only one bond to another atom. Carbon is unique among the elements in that it can form stable bonds with other carbon atoms to form long chains. Since a carbon in the middle of a chain has used only two of its bonds—one to bond to the carbon on its right, and the other to bond to the carbon on its left—it still has two more bonds to make. It can use one to bond, say, a nitrogen atom and the other perhaps to bond to another chain of carbon atoms.

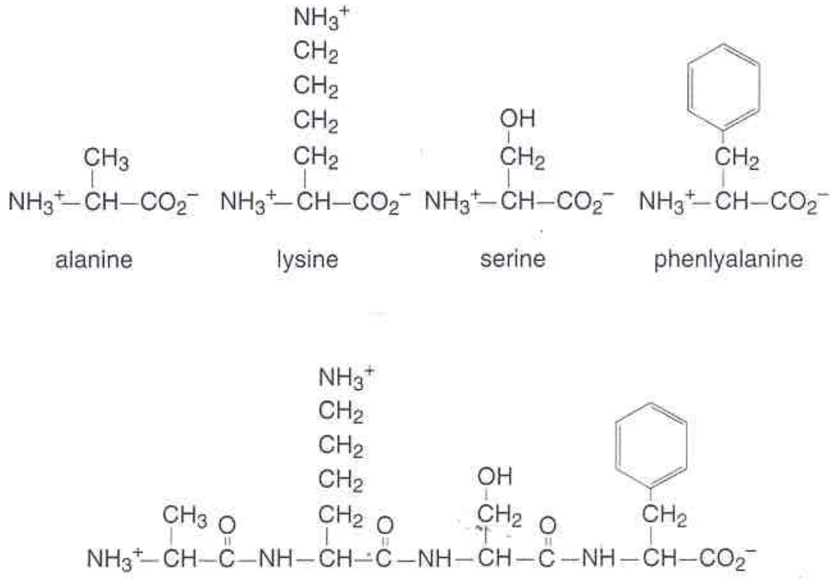
The number of molecules that can be built from carbon and the other biological elements is very large indeed. Biological systems, however, don't use a large number of completely different molecules. Rather, a limited number of molecules are made and the large, "macro" molecules of life—such as proteins, nucleic acids, and polysaccharides—are constructed by stringing together in different arrangements molecules from the limited set. This can be likened to making an enormous number of different words and sentences from the twenty-six letters of the alphabet.

The building blocks of proteins are called *amino acids*. The twenty different amino acids that compose virtually all proteins have a common structure. On the left side of the molecule is a nitrogen-containing group called an amine, and on the right, joined to the amine by a central carbon atom, is a carboxylic acid group (hence the name amino acid). Also attached to the central carbon, in addition to a hydrogen atom, is another group, called the side chain (Figure A–2). The side chain varies from one type of amino acid to another. It is the side chain that gives an amino acid its particular character.

Amino acids can be grouped into several categories. The first group contains hydrocarbon side chains (side chains with only carbon and hydrogen atoms). These side chains are oily, like gasoline, and prefer to avoid contact with water molecules. The next group is the electrically charged amino acids; there are three positively and two negatively charged members. Charged side chains prefer to be in contact with water. Another group is the polar amino acids. Polar molecules, although not fully charged, have partially charged atoms in them. This arises when one atom pulls more strongly on the electrons than its partner atom in a chemical bond, bringing the electrons closer to it. The atom with the lion's share of the electrons has a somewhat nega-

FIGURE A–2

(TOP) FOUR AMINO ACIDS. THE AMINO ACIDS DIFFER ONLY IN THEIR SIDE CHAINS. (BOTTOM) THE FOUR AMINO ACIDS HAVE BEEN CHEMICALLY JOINED. PROTEINS ARE LONG CHAINS OF MANY CHEMICALLY JOINED AMINO ACIDS.



tively charged character, while the atom with a deficiency of electrons has a partial positive charge. Interactions between positively and negatively charged side chains, and between the partially positively and partially negatively charged atoms of polar side chains, can be very important in the structure of proteins.

During the synthesis of proteins, two amino acids are chemically joined together by reacting the amino group of one amino acid with the carboxylic acid group of another to form a new group called a peptide bond (Figure A–2). The new molecule still has a free amino group at one end and a free carboxyl at the other end, so another amino acid can be joined by contributing its amino end to form another peptide bond. This process can be repeated indefinitely until a macromolecule, containing hundreds or thousands of amino acid "residues" (the part left after the chemical reaction joining two amino acids), has been formed. Such macromolecules are known as *polypeptides* or *proteins*.

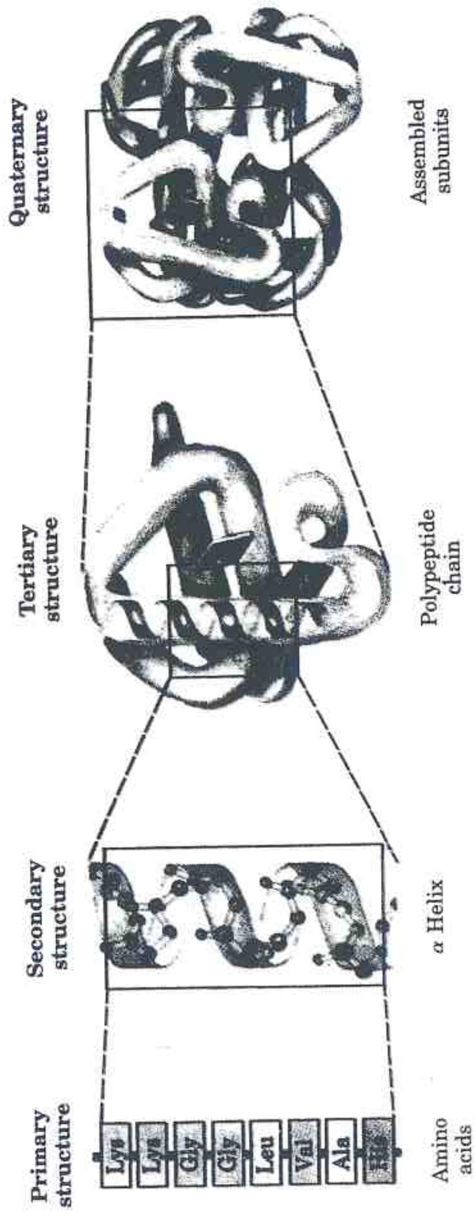
A typical protein contains anywhere from about fifty to about three thousand amino acid residues. The amino acid sequence of a protein is called its *primary structure*. The completed protein still has a free amino group at one end, referred to as the N-terminal end, and a free carboxyl at the other end, called the C-terminal end. The amino acid sequence of a protein is conventionally written starting from the N-terminal to the C-terminal end. The atoms of the protein joined in a line from the N to the C terminal are called the protein *backbone;* this includes all atoms except those of the side chains.

A freshly made protein does not float around like a floppy chain. In a remarkable process, virtually all biological proteins fold up into discrete and very precise structures (Figure A–3) that can be quite different for different proteins. This is done automatically through interactions such as a positively charged side chain attracting a negatively charged side chain, two hydrophobic side chains huddling together to squeeze out water, large side chains being excluded from small spaces, and so forth. At the end of the folding process, which typically takes anywhere from fractions of a second to a minute, two different proteins can be folded to structures as precise and different from each other as a three-eighths-inch wrench and a jigsaw. And, like the household tools, if their shapes are significantly warped, then they fail to do their jobs.

When proteins fold, they do not flop together like a string crushed

FIGURE A–3

THE FOUR LEVELS OF PROTEIN STRUCTURE.



Copyright Irving Geis. Reproduced with permission in your hand; there are regularities to the folding. Before a protein folds, its polar backbone atoms—the oxygen and nitrogen and hydrogen atoms in each peptide bond—form what are called *hydrogen bonds* to water molecules. A hydrogen bond occurs when a partially negatively charged peptide oxygen or nitrogen atom associates closely with the partially positively charged hydrogen atoms of water. When a protein folds, however, it must squeeze out all (or almost all) of the water so that the oily side chains can pack efficiently. This poses a problem: the polar peptide atoms must find oppositely charged partners in the folded protein, or else the protein will not fold.

There are two ways proteins solve this problem. First, segments of the protein can form an *α-helix*. In this structure the protein backbone spirals. The geometry of the spiral makes the oxygen atom of a peptide group point directly towards, and hydrogen bond with, the hydrogen of the peptide group found four amino acid residues back along the chain (Figure A–3). The next residue hydrogen bonds with the subsequent residue four back from it, and so on. Usually an α-helix has anywhere from five to twenty-five amino acid residues before the helical structure (but not necessarily the protein chain) ends. An α-helix permits a protein to fold into a compact shape while still forming hydrogen bonds to peptide atoms. A second structure that allows regular hydrogen bonding of peptide atoms is called a *ß-pleated sheet,* or simply a ß-sheet. In this structure the backbone of the protein goes up and down, like pleats in a sheet, and the peptide atoms stick out perpendicular to the direction of the protein chain. The chain then curls around, comes back, and the oxygen atoms in the peptide group of the returning strand hydrogen bond to the peptide group of the first strand. As with α-helices, ß-sheets allow polar backbone atoms to form hydrogen bonds.

α-helices and ß-sheets are known as the *secondary structure* of the protein. A typical protein has about 40 to 50 percent of its amino acid residues involved in helices and sheets. The remainder of the residues are involved in turns between portions of secondary structure, or else form irregular structures. Helices and sheets pack against each other to form, in most cases, a compact, globular protein. The exact way in which the elements of secondary structure pack is called the *tertiary structure* (Figure A–3) of the protein. The driving force for the packing of the helices and sheets comes from the oily nature of many protein side chains. Just as oil separates from water to form a distinct layer, so the oily, hydrophobic side chains huddle together to form a water-free zone in the interior of the protein. Recall, however, that some protein side chains are either polar or charged, and they want to stay in the water. The pattern of oily and polar side chains along the amino acid sequence, and the need for the protein chain to fold so that most of the hydrophobic groups are in the interior of the protein and most of the hydrophilic groups are on the exterior, provides the information that drives a specific protein to fold to a specific structure.

Another factor also contributes to the specificity of protein folding. In all folded proteins some polar side chains inevitably get buried. If the buried polar atoms do not find hydrogen-bonding partners, then the protein is destabilized. In most proteins about 90 percent of the buried polar side chain atoms are, in fact, hydrogen bonded to other side chains or to the protein backbone in a catch-as-catch-can manner. The folding of a typical protein—with its requirements to accommodate hydrophobic and hydrophilic groups and to form a network of hydrogen bonds—can be likened to a three-dimensional jigsaw puzzle.

Frequently, several separate polypeptides stick together in a very specific way to form a composite structure that functions as one entity. In these cases it is the custom to refer to the associated polypeptides as a single protein composed of several "subunits." For example, the oxygen-carrying protein hemoglobin is composed of four polypeptides, and the amalgamated protein has oxygen-binding properties that the component polypeptides lack. Thus the functional biological protein is the complex of the four polypeptides. The specific arrangement of separate polypeptides in a protein is called its *quaternary structure* (Figure A–3).

NUCLEIC ACID STRUCTURE

Like proteins, nucleic acids are polymers of a small number of building blocks, called *nucleotides.* A nucleotide itself has several parts. The first part is a carbohydrate, either ribose (in RNA) or deoxyribose (in DNA). To ribose is attached one of four *bases,* either adenine (A), cytidine (C), guanine (G), or uracil (U). If the carbohydrate is deoxyribose then U is replaced by a similar base called thymine (T); A, C, and G are also used with deoxyribose. Attached to a different part of the carbohydrate ring (to the 5'-OH or "five-prime hydroxyl" group) is a phosphate group. The sugar-phosphate portion of a nucleotide is analogous to the backbone portion of an amino acid, and the base is analogous to an amino acid side chain. It is only in its base that one nucleotide differs from another.

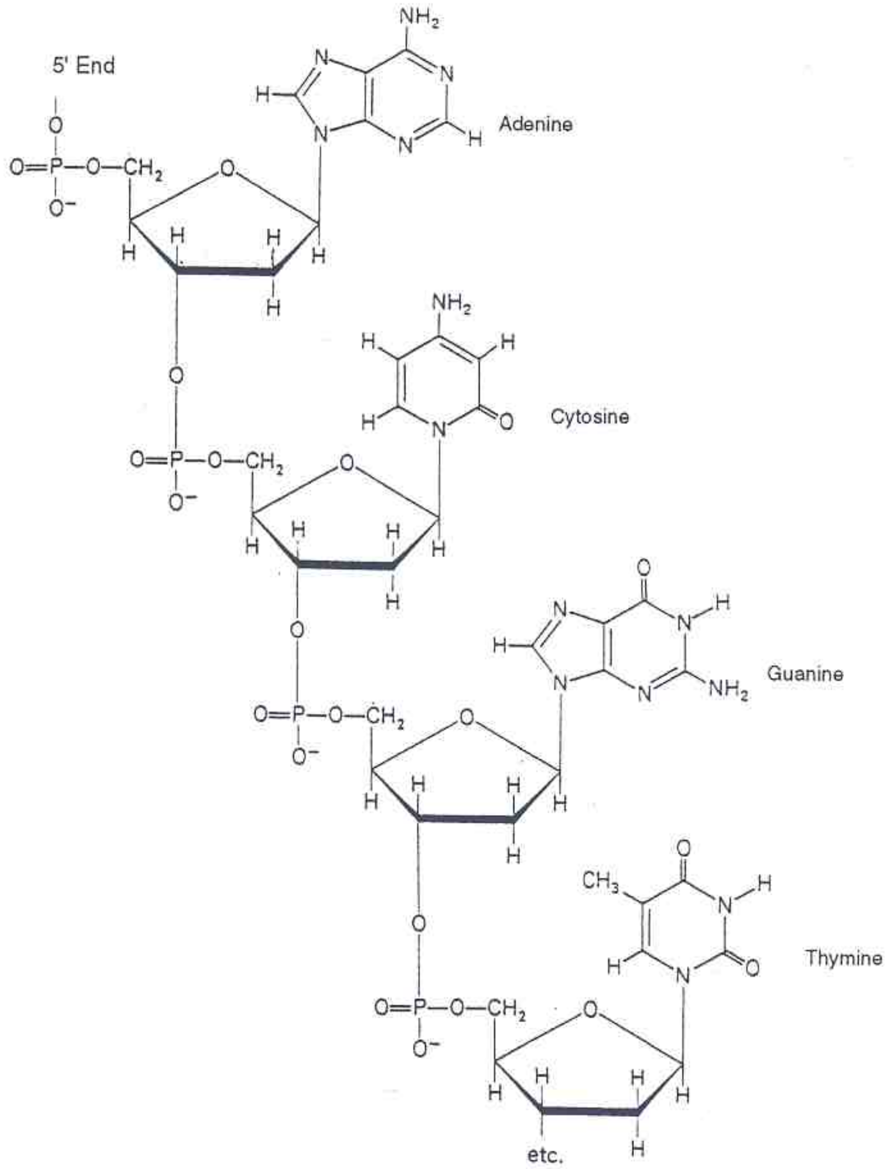
Two nucleotides can be joined chemically by reacting the phosphate of one nucleotide with the 3'-OH group of the carbohydrate portion of the second nucleotide (Figure A–4). This still leaves a free phosphate group on one end and a free 3'-OH group on the other end, which can be further reacted with other nucleotides. Repetition of this process can generate very long polynucleotides indeed. Cellular RNA ranges from about seventy to about fifty thousand nucleotides in length. One single molecule of DNA ranges from several thousand to about a billion nucleotides. The sequence of a polynucleotide is conventionally written starting from the 5' end to the 3' end.

Cellular RNAs are found as single polynucleotide chains. There are several biological classes of RNA. The first is called *messenger RNA* (mRNA); members of this class are produced as faithful transcripts of DNA genes; the genetic information carried by mRNA is then interpreted by the protein synthetic apparatus to produce a protein. The second type of RNA is called *ribosomal RNA* (rRNA). Polynucleotides in this class associate with a large number of different proteins to form the ribosome, the primary engine of protein synthesis. The last major category of RNA is called *transfer RNA* (tRNA). Members of this class are relatively small, seventy to ninety nucleotides in length, and serve as "adaptors" between the mRNA and the growing protein that is produced by the action of the ribosome.

Cellular DNA is found as a double-stranded molecule—two intertwined polynucleotides (the famous double helix) that are strongly held together by hydrogen bonding. To understand the reason for this we must look at the structure of the bases of the nucleotides (Figure A–4). The nucleotides can be divided into two categories—the purines (A and G), which carry the larger bases (composed of two fused rings), and the pyrimidines (C and T), which have only one ring. If A and T are correctly oriented, they can form two hydrogen bonds with each other, and G can form three hydrogen bonds with C. In cells, wherever there is a G in one strand of DNA there is a C in the second strand, and vice versa; and wherever there is an A in one strand there is a T in

FIGURE A–4

A PIECE OF DNA CONTAINING FOUR NUCLEOTIDES.



From Conn, E. E., Stumpf, P. K., Bruening, G., and Doi, R. H. (1987) *Outlines of Biochemistry,* 5th ed., John Wiley & Sons, New York, fig. 6.1. Reproduced with permission.

the second strand, and vice versa. Thus the two strands are called "complementary" to each other. To be correctly oriented for hydrogen bonding the two strands must be pointed in different directions, with one running 5' to 3' from left to right and the other going 5' to 3' from right to left. The DNA of eukaryotes consists of two complementary linear strands, but the DNA of many bacteria consists, surprisingly, of two complementary *circular* strands.

The amount of DNA in a cell varies roughly with the complexity of the organism. Bacteria have about several million nucleotides of DNA. The amount of eukaryotic DNA ranges from a low of several tens of millions of nucleotides in fungi to a high of several hundred billion in some flowering plants. Humans come in at around three billion nucleotides.

LIPIDS AND POLYSACCHARIDES

Two other major categories of biomolecules are lipids and polysaccharides. Polysaccharides are polymers of sugar molecules or their derivatives and play a variety of roles. They can be used as structural materials, such as the cellulose found in woody plants and trees, and as repositories of energy, such as the glycogen which is stored in the liver. Lipids, unlike proteins, nucleic acids, and polysaccharides, are not polymers made from discrete building blocks; rather, each lipid molecule must be synthesized from very basic starting materials. Lipids are not macromolecules, but they can associate to form large structures such as membranes.

TRANSCRIPTION

DNA, the repository of genetic information, is a polynucleotide. But the information it carries tells the cell how to make polypeptides—proteins. How does the information get translated from one polymer "language" to the other? Shortly after the discovery of the double helical structure of DNA physicist George Gamow proposed the very nonchemical idea that genetic information is stored in coded form, and that expressing the information involves decoding the polynucleotide and translating the message into the polypeptide language of proteins.[3] Although he was wrong about the specific nature of the code, Gamow's intuition was prophetic.

During the early 1960s the code was broken. Nobel laureates Marshall Nirenberg, Severo Ochoa, H. Gobind Khorana, and their associates showed that in the genetic code three contiguous nucleotides correspond to one amino acid (Figure A–5). Since there are sixty-four possible combinations of four bases taken three at a time, there are more than enough permutations to code for all twenty amino acids. All possible three base "codons" are used by the cell, so the genetic code is redundant, meaning that several different codons can designate the same amino acid. For example ACU, ACC, ACA, and ACG all code for the amino acid threonine. Most amino acids have two or more codons designating them; several, however, have only one. A total of sixty-one of the possible sixty-four codons designate amino acids; the remaining three are used as "stop" codons. When the decoding apparatus encounters one of these special signals, it halts its production of protein at that point.

The large number of steps involved in extracting the information contained in DNA can be divided into two conceptual categories called *transcription* and *translation*. Briefly, in transcription a cell makes an RNA copy of a small portion of its DNA (termed a *gene*) that

FIGURE A–5

THE GENETIC CODE.

UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine
UUC		UCC		UAC		UGC	
UUA	Leucine	UCA		UAA	Stop	UGA	Stop
UUG		UCG		UAG		UGG	Tryptophan
CUU		CCU	Proline	CAU	Histidine	CGU	Arginine
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	Glutamine	CGA	
CUG		CCG		CAG		CGG	
AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	Lysine	AGA	Arginine
AUG	Methionine	ACG		AAG		AGG	
GUU	Valine	GCU	Alanine	GAU	Aspartic acid	GGU	Glycine
GUC		GCC		GAC		GGC	
GUA		GCA		GAA	Glutamic acid	GGA	
GUG		GCG		GAG		GGG	

codes for a protein; in translation the information in the RNA is used to produce a protein.

The transcription of a gene entails a number of decisions, the first of which is where along the huge DNA chain to start. The beginning position is generally marked by several special DNA sequences, called "promoters." In prokaryotes a sequence of DNA nucleotides (usually TCTTGACAT) called the "–35 region" occurs about thirty-five nucleotides before a gene; another sequence (usually TATAAT) called the "Pribnow box" occurs five to ten base pairs prior to the transcription initiation site. In addition to similar signals, eukaryotes have DNA sequences called "enhancers" thousands of base pairs away from the transcription start site; enhancers can greatly affect the rate at which a gene is transcribed.

To begin transcription, in prokaryotes a multisubunit enzyme called *RNA polymerase* binds to DNA. RNA polymerase consists of five polypeptide chains. Initially the enzyme binds loosely, moving along the DNA like cars on a roller coaster until it finds the promoter region of a gene. When it does, one of the protein's subunits, called σ, recognizes the promoter DNA sequence. Right after RNA polymerase finds the promoter sequence σ floats away, its job finished. In the absence of σ, RNA polymerase binds quite tightly to the DNA and can no longer move freely. Now its work begins. The RNA polymerase "melts" about ten base pairs of DNA, separating the two polynucleotide strands from each other over that region. This is necessary so that the RNA chain that will be made can "read" the DNA template by hydrogen bonding to it. Now the polymerase binds the activated form of a ribonucleotide that is complementary to the first DNA base where transcription starts. Next it binds the second ribonucleotide, complementary to the second DNA base.

Once the first two correct ribonucleotides are matched to the template, the RNA polymerase chemically joins them. The polymerase then moves down one position along the DNA template, keeping the DNA strands separate as it goes. It matches the third position with its corresponding activated ribonucleotide, and joins that to the growing chain. These steps are repeated along the gene at a very high rate, moving at approximately twenty to fifty nucleotides per second.

Transcription causes a problem: the movement of the polymerase through the interwound, helical DNA causes the DNA ahead of the polymerase to become tightly overwound.[4] This would cause transcription to slow down or halt completely except that another protein, called *topoisomerase*, untangles the DNA. It does this by a complicated maneuver—cutting one strand of the tangled DNA, passing the uncut DNA strand through the cut strand, and then resealing the cut.

Transcription stops when the RNA polymerase runs into a special DNA sequence. In prokaryotes it is a palindromic[5] region containing about six or seven GC base pairs followed by a region of the same length rich in AT base pairs. Some, but not all, genes require an additional protein, called ρ, to make the polymerase fall off the DNA.

GENE REGULATION

A typical bacterial cell contains thousands of genes, and a typical mammalian cell contains tens of thousands. How does a cell know when to transcribe a gene, and how does it select a specific gene from the thousands available? The problem of "gene regulation" is a major focus of research. Many details have been uncovered, but much remains murky. One of the simplest examples of gene regulation is the regulation of the life cycle of bacteriophage λ. Bacteriophages—the prokaryotic analogs to viruses—are bits of DNA wrapped in a protein coat. In order to make copies of itself, a bacteriophage must find a suitable bacterial cell, attach itself to the cell, and inject its DNA into the host. The DNA from the phage is quite small, coding for only about fifty genes. This is not sufficient to make its own replication machinery so, cleverly, the phage hijacks the host's machinery. Thus the phage is a parasite, unable to provide completely for itself.

Sometimes when bacteriophage λ invades a cell, the cell makes so many copies of λ that it bursts. This is called the *lytic* cycle. At other times, however, λ inserts its own DNA into the bacterial DNA, making a single molecule from two. There the λ DNA can rest quietly, be replicated along with the rest of the bacterial DNA when the cell divides, and bide its time. This is called the *lysogenic* cycle. When the bacterium, perhaps many generations later, runs into trouble (by, say, encountering high doses of ultraviolet light), the λ DNA in the bacterial DNA switches to the lytic mode. Only now does the phage make thousands of copies of itself, bursting the cell and spilling out new bacteriophages.

What switches bacteriophage λ from the lysogenic to the lytic cycle? When bacteriophage DNA enters the cell, RNA polymerase binds to a bacteriophage λ transcription promoter. One of the first genes to be expressed is for an enzyme, called an "integrase," that chemically inserts the λ DNA into the bacterial DNA. The enzyme does this by cutting the circular λ DNA at a specific site that has a sequence similar to a site in the host DNA, which the integrase also cuts. This leaves both pieces of DNA with complementary, "sticky" ends that hydrogen bond to each other. The integration enzyme then joins the pieces of DNA.

Another λ gene codes for a protein called a "repressor." The repressor binds strongly to a sequence of λ DNA which RNA polymerase must bind to start the lytic cycle. When λ repressor is there, however, RNA polymerase cannot bind, so the lytic cycle is switched off. There are actually three binding sites for repressor—all in a row. Repressor binds the first site more strongly than the second site, and the second more strongly than the third. The third site overlaps the promoter for the gene that codes for the repressor itself. This arrangement allows the repressor to be synthesized continuously until the third site is filled, at which point synthesis stops. If the concentration of repressor falls to the point where it dissociates from the third site, then the repressor gene is again turned on.

By this mechanism λ repressor regulates its own production. In the presence of some chemicals, ultraviolet light, or other damaging agents, however, a gene for an enzyme that specifically destroys λ repressor is switched on. When the repressor is removed from the first site, the gene for a protein called Cro is activated. Cro protein binds strongly to the third λ repressor binding site, shutting it off forever, and launching the bacteriophage into the lytic cycle. All the genes necessary for making copies of the λ DNA and packaging them into protein coats are now transcribed.

The control of the life cycle of bacteriophage λ is one of the simplest examples of gene regulation. The regulation of other gene systems, especially in eukaryotes, can involve dozens of proteins. Nonetheless, it is thought that most genes are regulated by systems analogous to that of λ, with feedback controls and multiple factors conniving to decide whether a single gene should be turned on.

TRANSLATION

Once the messenger RNA has been produced, the task turns to translating the message into a protein. This process is best understood in prokaryotes.

The transcribed mRNA is bound by a particle called a ribosome. Ribosomes are huge complexes consisting of fifty-two separate proteins (of which several are present in multiple copies) and three pieces of RNA with lengths of 120, 1,542, and 2,904 nucleotides. The ribosome can be readily broken down into two large pieces, called the 30S subunit and the 50S subunit.[6] Incredibly, the ribosome is self-assembling. Experiments have shown that when ribosomes are separated into their components and then remixed, under the right conditions the components will spontaneously reform ribosomes.

The ribosome has a problem similar to that of RNA polymerase: the ribosome must find the point in the mRNA at which to begin translation. In prokaryotes the site is marked by a tract called the Shine-Dalgarno sequence, about ten nucleotides upstream from the initiation site. Initiation occurs at the first subsequent AUG sequence. (AUG codes for the amino acid methionine.) In eukaryotes, initiation usually begins simply at the first AUG sequence from the 5'-end of the mRNA.

Ribosomes cannot bind directly to mRNA by themselves; several other factors are required. In prokaryotes three proteins called *initiation factors*—labeled IF-1, IF-2, and IF-3—are necessary. To begin translation, IF-1 and IF-3 bind to the 30S ribosomal subunit. This complex then goes on to bind (1) to a previously-formed complex of a tRNA molecule carrying methionine and bound to IF-2, and (2) to the mRNA molecule at the initiation site. Next, the 50S ribosomal subunit binds to the growing complex, causing IF-1, IF-2, and IF-3 to fall off. In eukaryotes, translation initiation goes through similar steps, but the number of initiation factors can be as high as ten or more.

In the next step a second tRNA molecule, associated with a protein named elongation factor Tu (EF-Tu), comes in carrying the appropriate amino acid and binds to the ribosome. A peptide bond forms between the two amino acids held on the ribosome. The first tRNA molecule now has lost its amino acid, and the two covalently bonded amino acid residues are linked to the second tRNA. At this point the first tRNA dissociates from the ribosome, the second tRNA moves into the site on the ribosome previously occupied by the first tRNA, and the ribosome moves precisely three nucleotides down on the mRNA. This translocation process requires another protein called EF-G for some as-yet-unknown function.

These steps are repeated until the ribosome reaches a three-nucleotide sequence that corresponds to a stop codon. Another protein, called *release factor,* binds to the stop codon, preventing the ribosome from moving there. Additionally, the release factor changes the behavior of the ribosome. Instead of simply sitting on the mRNA waiting for the release factor to move, the ribosome cuts the completed polypeptide chain from the final tRNA molecule to which it is still attached, and the protein floats free into solution. The inactive ribosome then dissociates from the mRNA, floats away, and is free to begin another round of protein synthesis.

Other factors, too numerous to mention in this brief sketch, are also necessary for a functioning translation system. These include the enzymes that chemically place the correct amino acid onto the correct tRNA, various mechanisms to "proofread" the translation, and the role of chemical energy, in the form of the activated nucleotide GTP, at every stage of translation. Nonetheless, this outline may give the reader both an idea of the process by which genetic information is expressed and also an appreciation for the intricacies involved in that expression.

DNA REPLICATION

There comes a time in the life of every cell when it turns to thoughts of division. One major consideration in cell division is ensuring that the genetic information be copied and handed down uncorrupted; a great deal of effort is invested in that task.

In 1957 Arthur Kornberg demonstrated that a certain enzyme could polymerize the activated forms of deoxynucleotides into a new DNA molecule that was an exact copy of whatever "template" DNA Kornberg threw into the reaction mixture. He called the enzyme *DNA polymerase I* (Pol I). The scientific community was ecstatic about the find. Over the years, however, it has been shown that Pol I's primary role is not to synthesize DNA during cell division; rather, it is to repair DNA that has been damaged by exposure to ult .violet light, chemical mutagens, or other environmental insults. Two other DNA polymerases, Pol II and Pol III, were later discovered. The role of Pol II remains murky: mutant cells lacking the enzyme exhibit no observable defects. Pol III has been identified as the major enzyme involved in DNA replication in prokaryotes.

DNA polymerase III is actually a complex of seven different subunits, ranging in length from about 300 to about 1,100 amino acid residues. Only one of the subunits does the actual chemical joining of nucleotides; the other subunits are involved in critical accessory functions. For instance, the polymerizing subunit tends to fall off the template DNA after joining only ten to fifteen nucleotides. If this happened in the cell the polymerase would have to hop back on hundreds of thousands of times before replication was complete, slowing replication enormously. However, the complete Pol III—with all seven subunits—does not fall off until the entire template DNA (which can be more than a million base pairs long) is copied.

In addition to a polymerizing activity Pol III possesses, ironically, a 3'→5' nuclease activity. This means that it can degrade polymerized DNA into free nucleotides, starting at a free 3' end and working back toward the 5' end. Now, why would a polymerase also degrade DNA? It turns out that the nuclease activity of Pol III is very important in ensuring the accuracy of the copying procedure. Suppose that the wrong nucleotide became incorporated into the growing DNA chain. Pol III's nuclease function allows it to step back and remove the incorrect, mispaired nucleotide. Correctly paired nucleotides are resistant to the nuclease activity. This activity is called "proofreading"; without it, thousands of times more errors would creep in when DNA was copied.

DNA replication begins at a certain DNA sequence, known appropriately as an "origin of replication," and proceeds in both directions at once along the parent DNA. The first task to be tackled during replication, as for transcription, is the separation of the two parent DNA strands. This is the job of the *DnaA* protein. After the strands are separated two other proteins, called *DnaB* and *DnaC,* bind to the single strands. Two more proteins are recruited to the growing "bubble" of open DNA: *single strand binding protein* (SSB), which keeps the two parent DNA strands separated while the DNA is copied; and *gyrase,* which unknots the tangles that occur as the complex plows through double stranded DNA.

At this point DNA polymerase can begin synthesis. But several problems arise. DNA polymerase cannot start synthesizing by joining two nucleotides the same way that RNA polymerase starts transcription; the DNA enzyme can only add nucleotides to the end of a preexisting polynucleotide. Thus the cell employs another enzyme to make a short stretch of RNA on the exposed DNA template. This enzyme can begin RNA synthesis from two nucleotides. Once the RNA chain has gotten to be about ten nucleotides long, the DNA polymerase can then use the RNA as a "primer," adding deoxynucleotides to its end.

The second problem occurs as the replication "fork" opens up. The synthesis of one strand of new DNA can proceed without difficulty; this is the strand that the polymerase makes as it reads the template in a 3'→5' direction, making a new strand in a 5'→3' orientation, as all polymerases do. But how to synthesize the second strand? If done directly, the polymerase would have to read the template in a 5'→3' direction and thus synthesize the new strand in a 3'→5' direction. Although there is no theoretical reason why this could not occur, no known polymerase synthesizes in a 3'→5' direction. Instead, after a stretch of DNA has been opened up, an RNA primer is made near the fork and DNA synthesis proceeds backward, away from the replication fork, in a 5'→3' direction. Further synthesis on this "lagging" strand must wait until the replication fork opens up another stretch of DNA; another RNA primer must then be made, and DNA synthesis proceeds backward toward the previously synthesized fragment. The RNA primers must then be removed, the gaps filled in with DNA, and the ends of the DNA pieces "stitched together." This requires several more enzymes.

The above description of prokaryotic DNA replication has been pieced together by the enormous efforts of a large number of laboratories. The replication of eukaryotic DNA appears to be much more complex, and therefore much less is known about it.