Chapter 1

Some Molecular Biology

The purpose of this chapter is to provide a brief introduction to molecular biology, especially to DNA and protein sequences. Ideally, the reader has taken a beginning course in molecular biology or biochemistry and can go directly to Chapter 2. Introductory textbooks often exceed 1000 pages; here we just give a few basics. In later chapters we introduce more biological details for motivation.

One of the basic problems of biology is to understand inheritance. In 1865 Mendel gave an abstract, essentially mathematical model of inheritance in which the basic unit of inheritance was a *gene*. Although Mendel's work was forgotten until 1900, early in this century it was taken up again and underwent intense mathematical development. Still the nature of the gene was unknown. Only in 1944 was the gene known to be made of DNA; and, it was not until 1953 that James Watson and Francis Crick proposed the now famous double helical structure for DNA. The double helix gives a physical model for how one DNA molecule can divide and become two identical molecules. In their paper appears one of the most famous sentences of science: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." That copying mechanism is the basis of modern molecular genetics. In the model of Mendel the gene was abstract. The model of Watson and Crick describes the gene itself, providing the basis for a deeper understanding of inheritance.

The molecules of the cell are of two classes: large and small. The large molecules, known as macromolecules, are of three types: DNA, RNA, and proteins. These are the molecules of most interest to us and they are made by joining certain small molecules together in polymers. We next discuss some of the general properties of macromolecules, including how DNA is used to make RNA and proteins. Then we give some more details of the biological chemistry that are the basis of these properties.

1.1 DNA and Proteins

DNA is the basis of heredity and it is a polymer, made up of small molecules called nucleotides. These nucleotides are four in number and can be distinguished by the four bases: adenine (A), ctyosine (C), guanine (G), and thymine (T). For our purposes a DNA molecule is a word over this four letter alphabet, $\mathcal{A} = \{A,C,G,T\}$. DNA is a nucleic acid and there is one other nucleic acid in the cell, RNA. RNA is a word over another four letter alphabet of ribonucleotides, $\mathcal{A} = \{A,C,G,U\}$ where thymine is replaced by uracil. These molecules have a distinguishable direction, and for reasons detailed later, one end (usually the left) is labeled 5′ and the other 3′.

Proteins are also polymers and here the word is over an alphabet of 20 amino acids. See Table 1.1 for a list of the amino acids and their one and three letter abbreviations. Proteins also have directionality.

How much DNA does an organism need to function? We can only answer a simpler question: How much DNA does an organism have? The intestinal bacterium *Escherichia coli* (*E. coli*) is an organism with one cell and has about 5×10^5 letters per cell. The DNA contained in the cell is known as the *genome*. In contrast to the simpler *E. coli*, the genome of a human is about 3×10^9 letters. Each human cell contains the same DNA.

Both RNA and proteins are made from instructions in the DNA, and new DNA molecules are made from copying existing DNA molecules. These processes are discussed next.

1.1.1 The Double Helix

The key feature of DNA that suggested the copying mechanism is the complementary basepairs; that is, the bases pair with A pairing T and G pairing C. This so-called pairing is by hydrogen bonds; more on that later. The idea is that a single word (or strand) of DNA (written in the 5′ to 3′ direction)

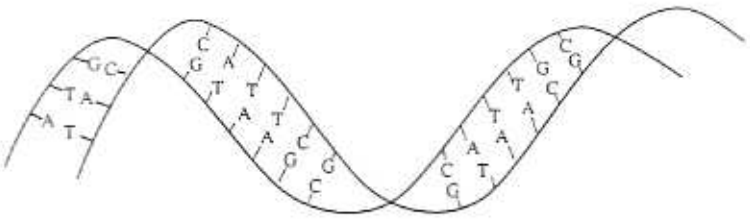
5′ACCTGAC3′

is paired to a complementary strand running in the opposite direction:

```
5'ACCTGAC3'
  | | | | | | |
3'TGGACTG5'
```

There are seven basepairs in this illustration. The A-T and G-C pairs are formed by hydrogen bonds, here indicated by a heavy bar. DNA usually occurs double stranded and its length is often measured by number of basepairs.

The three-dimensional structure is helical. In the next figure we show the letters or bases as attached to a string or backbone; note that the bars indicating the hydrogen bonds have been deleted. To properly view this figure, imagine a ribbon with edges corresponding to the backbones twisted into a helix.

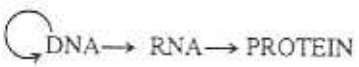


1.2 The Central Dogma

DNA carries the genetic material–the information required by an organism to function. (There are exceptions in the case of certain viruses where the genetic material is RNA.) DNA is also the means by which organisms transfer genetic information to their descendants. In organisms with a nucleus (eukaryotes), DNA remains in the nucleus; whereas proteins are made in the cytoplasm outside of the nucleus. The intermediate molecule carrying the information out of the nucleus is RNA. The information flow in biology is summarized by the "central dogma," put forward by Francis Crick in 1958:

> The central dogma states that once 'information' has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but transfer from protein to protein, or from protein to nucleic acid, is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.

A schematic for the central dogma is:



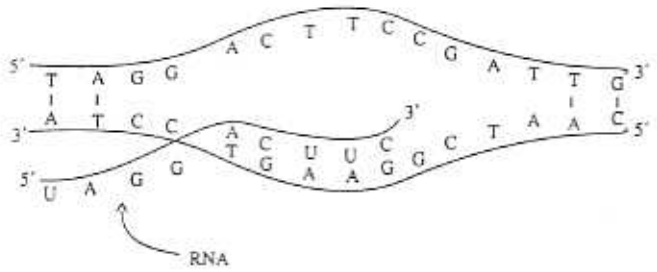
The loop from DNA to DNA means that the molecule can be copied. This is called replication. The next arrow is called transcription and the last translation. This chapter explores the arrows of the schematic in more detail.

Each of the arrows indicates making another macromolecule guided by the sequence of an existing macromolecule. The general idea is that one macromolecule can be used as a template to construct another. The fascinating details of these processes are basic to life. Understanding of templating will give us insight into the reasons for some interesting analytical studies. Today the central dogma has been extended. There are examples of genetic systems in which RNA templates RNA. Also, retroviruses can copy their RNA genomes into DNA by a mechanism called reverse transcription.

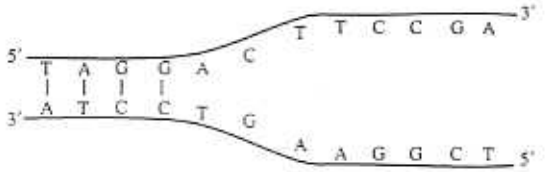
Making new molecules is called *synthesis*. When we look in detail we will see that certain proteins are required for the synthesis of both RNA and DNA. In

other words, we are about to sketch a highly complex system. For now it is easy to see how DNA can be a template to make new DNA.

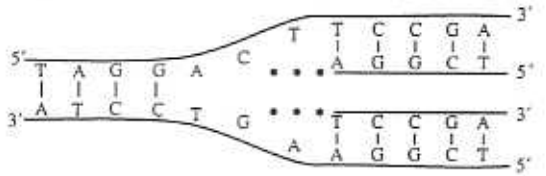
RNA is made single stranded. First the strands of the double helix are separated in a region by breaking the hydrogen bonds forming the basepairs. One strand of the DNA is used to template a single strand of RNA that is made by moving along the DNA. At the conclusion, the double stranded DNA remains as before and a single strand of RNA has been made. In the next illustration, the RNA is made from 5′ to 3′. Note that where T's existed in the complementary DNA, U's exist in the complementary RNA.



Making a new DNA from one already existing is called DNA *replication*. We begin with a double helix that has been separated into two single strands.



Then the single strands are used to template new double strands.



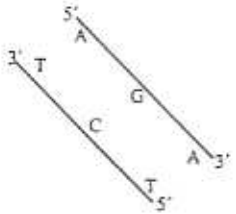
In this way two identical DNA molecules are made, each having one strand of the original molecule. Replication in this picture proceeds from right to left.

1.3 The Genetic Code

As soon as Watson and Crick proposed the double helix model of DNA in 1953, scientists began to study the problem of how a linear or helical DNA molecule

could encode a linear protein molecule. Cracking the genetic code became a hot topic and even attracted George Gamow (of the Big Bang Theory), a physicist. The sequence of insulin was the only protein sequence available and it was scrutinized very carefully. At that time it was not known that all amino acid sequences could be encoded in genes. Gamow, concentrating on the insulin sequence and the fact that 20 amino acids are used in protein sequences, discovered a very compelling code.

By example, consider the helix



Gamow extracted the diamond

and, reasoning that the code should be the same in either direction, he decided that

should encode the same amino acid. The second diamond is obtained by rotating the first one by 180°. Let us count the diamonds which encode amino acids in this scheme. There are two basepairs A · T and G · C. The number of diamonds with top and bottom base identical is, therefore, $\binom{4}{1} \times 2 = 8$. Otherwise, the top and bottom bases are unequal and the number of diamonds is $\binom{4}{2} \times 2 = 12$. In this scheme, the orientation of the basepair is not counted. When Gamow realized this he concluded that $20 = 8 + 12$, and he had found a candidate for the genetic code. The restrictions imposed on the possible sequences of amino acids by Gamow's scheme were severe, and his idea was rejected even before the genetic code was solved.

Crick's approach was to assume the code reads blocks of letters. These blocks cannot be less than 3 letters long: 4 and 4^2 are both less than 20, whereas $4^3 = 64$ exceeds 20. He probably came to this approach by reasoning that each strand was a template for the double helix, so should be sufficient to code proteins too. Crick decided that the genetic code should be "comma-free" – that the reading frame is determined by the blocks. Thus, if amino acids are encoded by triplets of nucleotides of DNA (*codons*) and if the code is comma-free, the reading frame of three consecutive nucleotides is

$$\underbrace{x_1x_2x_3}_{R_1}\ \underbrace{x_4x_5x_6}_{R_2}\ \underbrace{x_7x_8x_9}_{R_3}\ \cdots,$$

and not

$$x_1\ \underbrace{x_2x_3x_4}\ \underbrace{x_5x_6x_7}\ \underbrace{x_8x_9x_{10}}$$

or

$$x_1x_2\ \underbrace{x_3x_4x_5}\ \underbrace{x_6x_7x_8}\ \underbrace{x_9x_{10}x_{11}}\ ,$$

so that there is only one reading frame encoding $R_1R_2R_3\ldots$.

Once again the magic number 20 comes out of counting. The assumption or requirement is that all possible amino acid sequences are possible. Clearly AAA, TTT, GGG, and CCC are all impossible because in AAAAAA there is no obvious reading frame. (There are four places to begin reading AAA.) Therefore if $4^3 = 64$ possible codons are being considered, we are left with $4^3 - 4 = 60$ to study.

Of those remaining, let XYZ be the codon. Clearly, to have a comma-free code, XYZXYZ must be read unambiguously, and whenever XYZ is a codon, YZX and ZXY are not. The number of remaining codons equals $1/3 \times 60 = 20$. Alas, it turns out that biology has found a different and less mathematically elegant solution.

The genetic code can be read from a single strand of RNA, and it is read 5′ to 3′. The code is a triplet code: nonoverlapping successive blocks of three letters are translated into amino acids. There is a defined start or reading frame. Table 1.2 gives the genetic code in a compact form. There are three triplets–codons–that cause protein transcription to cease: UAA, UAG, and UGA. Viewed abstractly, the genetic code is a language in which 64 possible combinations of the 4 bases–uracil (U), cytosine (C), adenine (A), and guanine (G)–taken 3 at a time specify either a single amino acid or termination of the protein sequence. With 64 possible "words" and 21 possible "meanings," there is clearly the potential for different codons coding for identical amino acids. These 21 meanings are the 20 amino acids plus termination or stop. This is, in fact, the case: many pairs of codons that differ only in the third position base code for the same amino acid. On the other hand, a pair of codons differing only in the first or second position usually code for different amino acids.

RNA that is translated into protein is known as messenger RNA, mRNA. For example

$$\begin{array}{ll} \text{mRNA} & \underbrace{\text{UUUUACUGCGGCC}}\cdots \\ \text{protein} & \text{Phe Tyr Cys Gly} \quad \cdots \end{array}$$

A shift of one letter in reading the same nucleic acid sequence results in a very different amino acid sequence:

$$\begin{array}{ll} \text{mRNA} & \text{U}\,\underbrace{\text{UUUACUGCGGCC}}\cdots \\ \text{protein} & \cdots\ \text{Phe Thr Ala Ala}\ \cdots \end{array}$$

amino acid	3 letter code	1 letter code
alanine	Ala	A
arginine	Arg	R
aspartic acid	Asp	D
asparginine	Asn	N
cysteine	Cys	C
glutamic acid	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
methionine	Met	M
phenylalanine	Phe	F
proline	Pro	P
serine	Ser	S
threonine	Thr	T
tryptophan	Trp	W
tyrosine	Tyr	Y
valine	Val	V

Table 1.1: *Amino acid abbreviations*

The phase of codon reading is called the reading frame. There are three reading frames going 5′ to 3′. Reading the complementary DNA strand, there are three reading frames in the opposite direction. Therefore, there are a total of six possible reading frames possible for double stranded DNA.

The genetic code was solved in a very interesting way. Extracts from bacteria were prepared except for the RNA template, mRNA. The extracts would then make a protein sequence when the experimenter added synthetic mRNA. They began by adding UUUUU$\cdots$ and the proteins synthesized were composed of phenylalanine Phe - Phe -, ... Because the reading frame is irrelevant, this suggests UUU codes for Phe. Next, try a mRNA such as UGUGUG..., where the results are polypeptides of Cys or Val. This still does not allow us to assign a unique codon so we try UUGUUG$\cdots$, where the resulting polypeptides are made of Leu, Cys, and Val. This does not yield a codon assignment even though {UGU, GUG} $\cap$ {UUG, UGU, GUU}={UGU}. Next try UGGUGG$\cdots$ where the peptides contain Trp, Gly, and Val. The common codon is {UGU, GUG} $\cap$ {UGG,GGU,GUG} = {GUG} so we assign GUG to Val, the common amino acid. Most of the genetic code was cracked in this manner.

Let N = {A, C, G, U} be the set of nucleic acids, C = $\{(x_1x_2x_3) : x_i \in \mathrm{N}\}$,

Table 1.2: *The genetic code (64 triplets and corresponding amino acids) shown in its most common representation. The three codons marked* TC *are termination signals of the polypeptide chain.*

	2nd				
	U	C	A	G	
1st					3rd
	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	TC	TC	A
	Leu	Ser	TC	Trp	G
	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
A	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

and A be the set of amino acids and termination codon. The genetic code is simply a map $g : \mathbf{C} \rightarrow \mathbf{A}$.

1.4 Transfer RNA and Protein Sequences

As we have mentioned above, mRNA is read to make proteins. The amino acids are made available in the cell, and some are synthesized by the cell itself. With the amino acids and mRNA in the cell, there is an obvious question: How does a protein get made?

Part of the answer lies with the so-called adapter molecule, another RNA molecule known as transfer RNA (tRNA). Amino acids are linked to these smaller tRNA molecules of about 80 bases, and the tRNA then interacts with the codon of the mRNA. In this way, tRNA carries the appropriate amino acids to the mRNA. Obviously these reactions must be very specific. To understand this process it is

necessary to closely examine tRNA.

As RNA is single stranded, without the complimentary strand that DNA has, the molecule tends to fold back on itself to form helical regions. For example, 5′ GGGGAAAAACCCC 3′ can form the structure

```
G G G G A A
| | | |     A
C C C C A A
```

with a helix of 4 GC basepairs. This structure is known as a hairpin with a four basepair stem and a five base loop. Later in Chapter 13 we will study prediction of

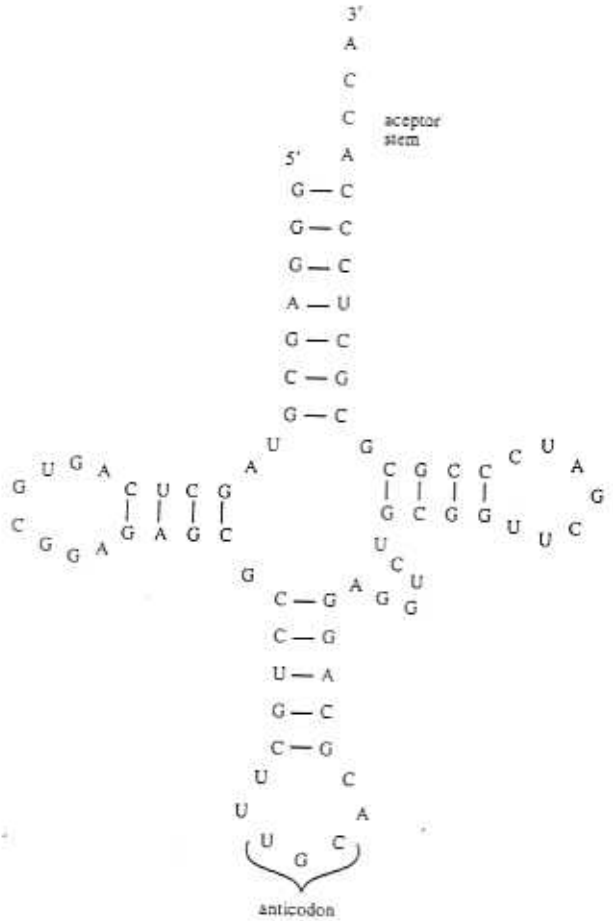


Figure 1.1: E. coli *Ala tRNA*

RNA structure. The longer sequence of a tRNA forms a more complex structure known as a cloverleaf. See Figure 1.1 for an *E. coli* tRNA associated with Ala. Next, the cloverleaf structure is given schematically, where only the backbone is shown.

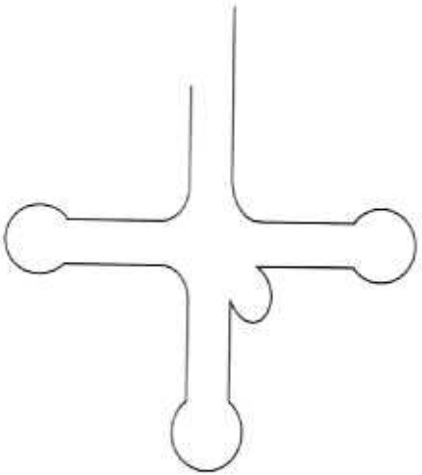
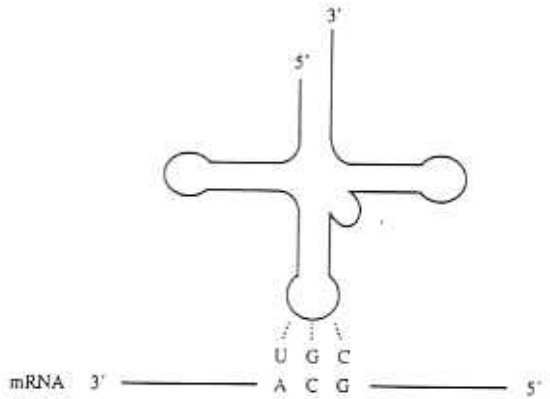


Figure 1.2: *tRNA schematic*

Actually this schematic only shows the simplest components of tRNA structure. There are some additional bonds formed and the entire structure becomes L shaped, with the 3′ ACCA sequence at one end and the anticodon at the other. As the name indicates, an *anticodon* is complementary to the codon, and three basepairs can form between the codon in the mRNA and the anticodon of the tRNA. For example, for the Ala codon GCA, we have the following



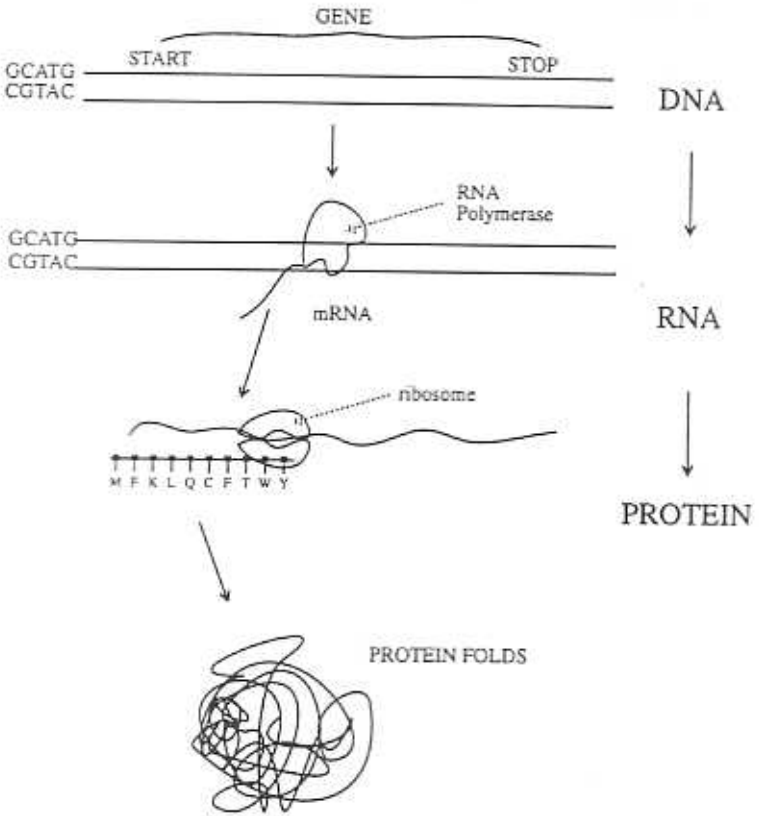


Figure 1.3: *DNA to RNA to protein*

Given that the codon GCA encodes Ala, the interaction between the tRNA and the mRNA seems almost inevitably to involve basepairing. The triplet UGC is the anticodon.

At the ribosome, mRNA is read and tRNA is utilized to make the protein sequence. From DNA to RNA to linear protein sequence to folded protein is shown in Figure 1.3.

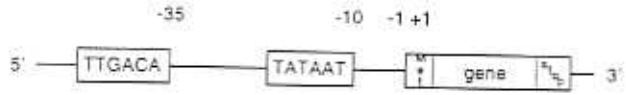
1.5 Genes Are Not Simple

In this section we will look at some complexities and variations of how genes work in living systems. It is not possible to do more than briefly mention some of these fascinating topics, but the reference books contain much fuller treatments.

1.5.1 Starting and Stopping

In the genetic code (Table 1.2) there are three codons for "stop," signaling the end of the gene. Not mentioned there is the fact that genes begin with the so-called start codon, AUG, which codes for Met. As is often the case in biology, the story is not so simple, and the details are highly dependent on the organism. In this section we will describe a well-studied system, the bacterium *Escherichia coli* or *E. coli*.

A molecular complex of several proteins called RNA polymerase is required to transcribe mRNA from DNA. For reasons of efficiency and control, there are signals in DNA to start and stop RNA transcription. The canonical start pattern has specific small sequences in the DNA as follows:



The idea is that the polymerase binds to these two patterns and then is in position to proceed down the DNA, transcribing it into RNA. The sequences the polymerase binds to are called *promoter sequences*. The Met initiator codon is 10 or so bases beyond the mRNA start at +1. (There is no 0 in the numbering schemes of biological sequences.) These patterns are not precise in content or in location. Later in Chapter 10 we will study ways to discover these patterns in bacterial promoter sequences.

1.5.2 Control of Gene Expression

Proteins from different genes exist in widely varying amounts–sometimes in ratios of 1/1000. Gene expression could be controlled at two points: DNA → RNA or RNA → protein. One common way of regulating a gene is by a repressor, which affects the step DNA → RNA. Suppose that the gene exists to process a molecule such as the sugar lactose. When lactose is absent, a repressor molecule (another protein) binds the DNA, stopping the DNA → RNA step. When lactose is present, it binds to the repressor and, in turn, prevents it from binding DNA. Of course, when the expressed gene (the protein) has processed all the lactose molecules, the repressor is no longer inhibited by lactose and the repressor again binds DNA, shutting down the transcription of the gene.

This clever scheme allows the organism to only make protein to process lactose when needed, thereby saving much unneeded RNA and protein. This

simple device is just one of a long and complex series of control mechanisms the cell has invented to deal with various environmental situations and various developmental stages.

1.5.3 Split Genes

Initially sequencing was done for *E. coli*, a member of the *prokaryotes*, organisms without a nucleus. When more rapid DNA sequencing began in 1976-1977, reading the genes of *eukaryotes*, organisms with a nucleus, was an obvious goal. There was soon a great surprise: The DNA encoding proteins was interrupted by noncoding DNA that somehow disappeared in the mRNA. Biologists, for example, expected $E_1E_2E_3$ to appear as one continuous coding region. Instead I_1 and I_2 split the gene into two pieces.

The so-called *exons*, E_1, E_2, and E_3, become an uninterrupted sequence, whereas the so-called *introns* I_1 and I_2 are spliced out and discarded. See Figure 1.4. A gene of 600 bases might be spread out over 10,000 bases of DNA. In yeast there is a tRNA gene that has 76 bps interrupted by a 14-bp intron. In a human gene, thyroglobin, 8500 bps are interrupted by over 40 introns of 100,000 bps.

Much remains to be learned about introns and exons. Why did they evolve? How can we recognize genes in uninterpreted DNA? What are the signals for splicing out the introns? These interesting questions do not yet have simple answers although much has been learned.

Originally it was supposed that most of the DNA encoded genes. It turns out to be true for viruses where it is important to be compact. In higher organisms, this is far from the case. Humans have around 5% of the genome used in protein coding. The function of much of the remaining DNA is unknown. Many people feel that much of it is "junk DNA," just sitting around not used for anything; others think that this DNA has important biological functions that are not yet understood.

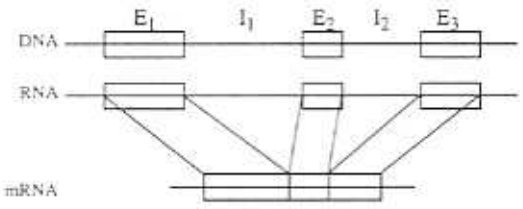


Figure 1.4: *Exons and introns*

1.5.4 Jumping Genes

One idea of molecular evolution is that it proceeds in small local steps. Our concept of a genome is that it is a blueprint for an organism. This concept is significantly altered by the discovery in both prokaryotic and eukaryotic genomes of segments of sequence that move from place to place in the genome. These sequences are known as transposable elements. They carry genes required for their movement or transposition, hence the name "jumping genes".

Much speculation has been made regarding the role of transposable elements. They of course can carry genetic material into new locations in the genome. In addition, as they often propagate themselves, they create identical or similar segments of DNA in various places in the genome. This can set the stage for duplication or deletion of the DNA between the transposable elements.

The role of transposable elements is not very clear. Some have suggested that transposable elements are "selfish DNA" and exist only for their own well being. That is, these elements could be viewed as mini-organisms themselves, living in the larger environment of genomic DNA.

In case this story sounds like an oddity, isolated to some obscure organisms, we point out that all organisms examined for transposable elements have been found to have them. Bacteria to humans, we all have jumping genes.

1.6 Biological Chemistry

It is now commonly understood that the molecules of biological organisms obey the standard and familiar laws of chemistry and physics. Until very recently it was thought otherwise, that there was a special set of laws of nature that apply to living organisms–perhaps a vital force. In fact, the chemistry of living organisms is special due to the requirements of organization and replication. In this chapter we will briefly touch on some of the basics of this chemistry. As pleasant as it is for mathematical scientists to view a DNA molecule as a long word over a four letter alphabet, it is very helpful to understand a little more of the basics.

The Basic Atoms

We will refer to the molecules of living organisms as biomolecules. Most biomolecules are made of only six different atoms: carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur. Table 1.3 shows some properties of the atoms. These atoms combine into the fantastic variety of organisms by means of chemical bonds.

The most abundant elements in living organisms are carbon, hydrogen, nitrogen, and oxygen. The are found in all organisms. In addition, calcium (Ca),

Atom	# electrons in outer shell	usual # covalent bonds
carbon (C)	4	4
hydrogen (H)	1	1
nitrogen (N)	5	3,5
oxygen (O)	6	2
phosphorus (P)	5	3,5
sulfur (S)	6	2 (up to 6)

Table 1.3: *Covalent bonds*

chlorine (Cl), magnesium (Mg), phosphorus, sodium (Na), and sulphur are present in all organisms, but in much smaller amounts. Finally, cobalt (Co), copper (Cu), iron (Fe), manganese (Mn), and zinc (Zn) are present in small amounts and are essential for life. Other elements are utilized by some organisms in trace amounts.

Covalent Bonds

Recall that a covalent bond is formed when two atoms are held together because electrons in their outer shells are shared by both atoms. Covalent bonds are the strongest of the variety of bonds between biomolecules and contribute to great stability. The outer, unpaired electrons are the only ones that participate in covalent bonds. There cannot be more covalent bonds than electrons in the outer shell, but all outer shell electrons, by no means, need be used in covalent bonds.

The arrangement of the atoms in space is only hinted at in the chemical structure. The ball and stick model of methane has the hydrogen atoms at the points of a tetrahedron. Water approximates that structure due to two groups of two electrons that occupy the points with the two hydrogen atoms. Another complication of the water molecule comes from unequal sharing of electrons in the covalent bonds; such bonds are called dipolar.

The stability of the covalent bond can be quantized by the potential energy of the bond. The energy is given in the number of kilocalories in the bonds of a mole (6.02×10^{23} molecules), the units are kcal/mol. In these units an O–H has 110 kcal/mol, C–O has 84 kcal/mol, S–S has 51 kcal/mol, and C=O has 170 kcal/mol. The range is approximately 50 to 200. Most chemical bonds found in biological molecules are 100 kcal/mol.

Weak Bonds

We will discover that the structure or three-dimensional shape of a molecule is extremely important in biology. Often these structures as well as molecular interactions are stabilized by bonds much weaker than those discussed above. The energies of these weaker bonds are in the range of 1 to 5 kcal/mol. They

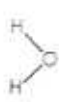
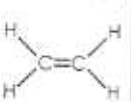
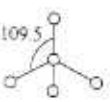
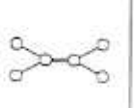
Name	water	methane	ethylene
Shorthand	H_2O	CH_4	H_4C_2
Chemical Structure			
Ball and Stick Model			
Space Filling Model			

Table 1.4: *Molecular models*

exist in a range that allows them to be formed or broken easily. This is because the kinetic energy of molecules at physiological temperatures ($\approx 25°C$) is about 0.5 kcal/mol. Just a few of these weak bonds can stabilize a structure, but the structure can then be altered as necessary.

There are several types of weak bonds.

• **The hydrogen bond** is a weak electrostatic bond forming between an negatively charged atom (frequently oxygen) and a hydrogen atom that is already covalently bound. The hydrogen atom is positively charged. Two such bonds are

$$C = O \cdots H - N,$$

$$C = O \cdots H - O$$

The strengh of a hydrogen bond is from 3 to 6 kcal/mol.

• **The ionic bond** is formed between oppositely charged components of molecules. These bonds would often be very strong if they were not in water, which reacts with the components decreasing the bonding energy. In solids the bond could be 80 kcal/mol, whereas in solution it might be 1 kcal/mol.

• **Van der Waals interactions** occur when two atoms are close together. Random

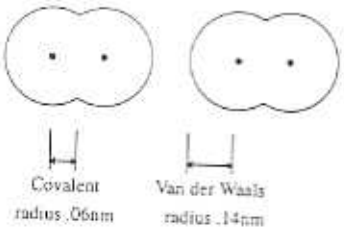


Figure 1.5: *Covalent bonds and Van der Waals interactions.*

movements of the electrons create dipole moments in one atom which generate an attractive dipole in the second. This causes a weak interaction that occurs between all molecules, polar or not. These interactions are nonspecific; that is, they do not depend on the specific identity of the molecules.

Figure 1.5 is a sketch of two O_2 molecules in Van der Waals contact. Van der Waals bonds have between 0.5 and 1.0 kcal/mol.

• **Hydrophobic interactions** occur among nonpolar molecules that cannot interact with water. The attraction is due to their aggregation in water, due to the noninteraction with water. Hydrophobic bond strength is between 0.5 and 3.0 kcal/mol.

Classes of Biomolecules

Many small molecules are present in organisms. They are needed for various reactions or are the product of these reactions. The general classes of small molecules are sugars, fatty acids, amino acids, and nucleotides.

Large molecules are built from the small molecules. The large molecule/small molecules relationships are polysaccharides/sugars, lipids/fatty acids, proteins/a--mino acids, nucleic acids/nucleotides. The last two large molecules, proteins and nucleic acids, are so big that they are known as macromolecules. This book is largely the mathematical study of these molecules and their biological properties.

Life is complex and the complexity requires these macromolecules. The amount of DNA required for life varies with the organism. Humans have 46 chromosomes, each if untangled and extended, about 4cm in length. Therefore the entire DNA in a nucleus of a single cell is about 2 meters in total length. The reason for this large size is the need to include all the information required to encode a human. While the human genetic material or genome is about 1000 times that of a bacterium, even bacterial genomes are large. Generally the size of the genome is an indicator of the size of the organism, but this is not a fixed rule.

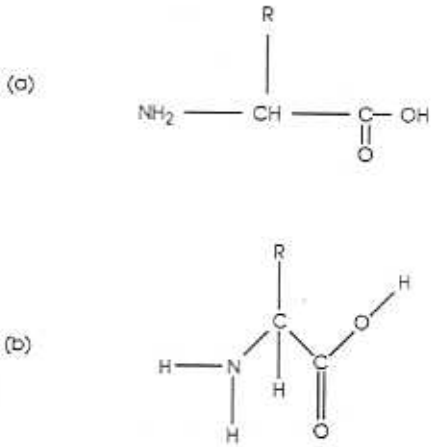


Figure 1.6: *General chemical structure of an amino acid (a) with more detail shown in (b)*

For example, the genomes of certain lilies and lungfish are about 100 times larger than the human.

Proteins

Proteins are the structural elements and the enzymatic elements of an organism–they are the working parts as well as the building material. These very important macromolecules are made of a sequence of molecules called amino acids.

There are 20 amino acids with the general chemical structures shown in Figure 1.6. The "R" in Figure 1.6 stands for the variable element or group which is known as a side chain, R group, or residue R. R gives the amino acid its identity; there are 20 R's and, consequently, 20 amino acids. COOH is known as the carboxyl group and NH_2 as the amino group. The central carbon is known as the α carbon. This atom is often used to locate an amino acid in a protein.

How do amino acids become proteins? There are many levels at which to approach this question. Here we focus on the most elementary chemical view and see in Figure 1.7 how three amino acids with residues R_1, R_2, and R_3 can be joined to form a three residue protein $R_1R_2R_3$ plus two water molecules.

Note that there is a chemically defined direction or orientation to this molecule. It "begins" with the amino end (N) and proceeds to the carboxyl end (C). The number of unique proteins is enormous, as there are 20^n proteins of length n.

The 20 amino acids structures are shown in Figure 1.8 with some of their properties.

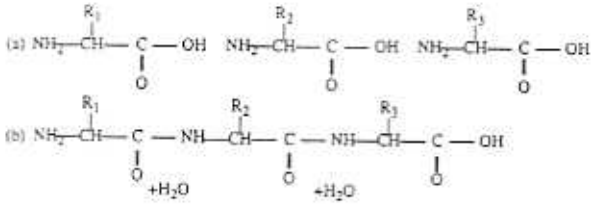


Figure 1.7: *(a) Three amino acids with residues* R_1, R_2, *and* R_3. *(b) A protein or polypeptide with residues* $R_1R_2R_3$, *plus the two water molecules.*

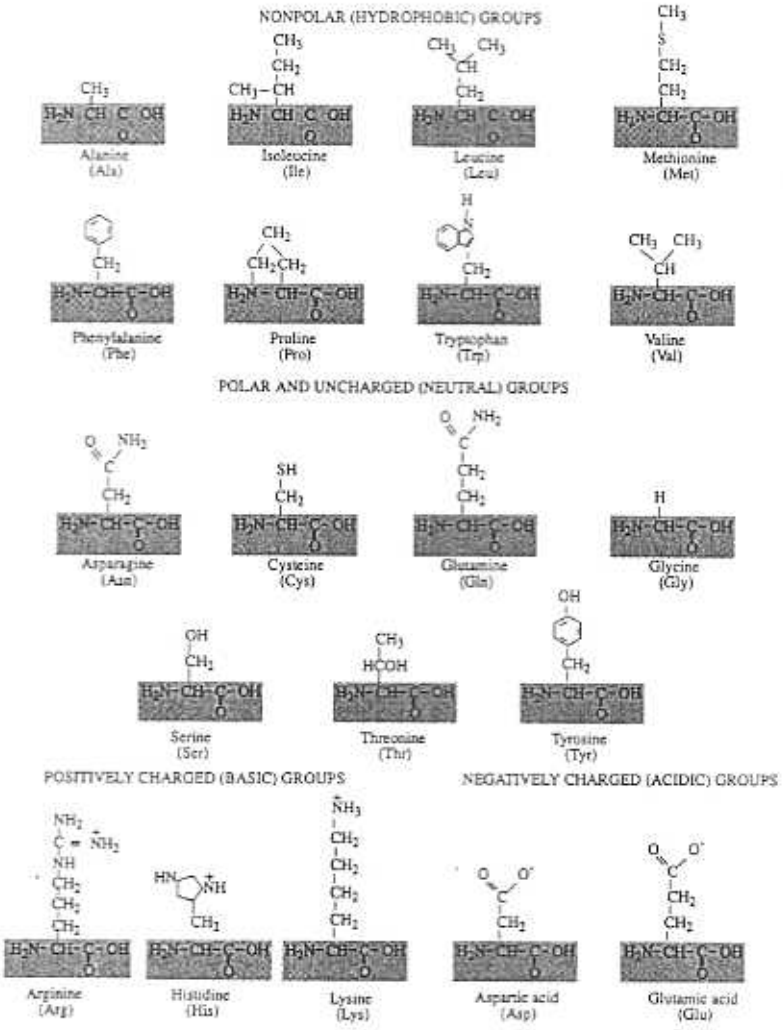
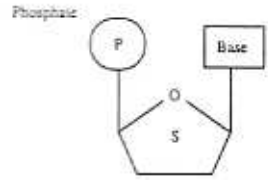


Figure 1.8: *Chemical structure of the 20 amino acids*

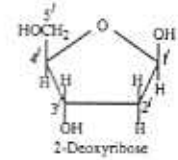
DNA

Deoxyribonucleic acid (DNA) is the carrier of genetic information for all organisms, except for some viruses. DNA consists of four different units called nucleotides. The components of a nucleotide are a phosphate group, a pentose (a 5-carbon sugar) and an organic base. The four different bases determine the identity of a nucleotide.

The general picture of a nucleotide is



In more detail, the structure of the sugar S (a pentose) is



The carbon atoms are often not shown in the figure; only one appears in the above sugar structure. The numbers $1'$ to $5'$ refer to the carbon atom locations in the sugar. The carbon atoms $5'$ and $3'$ are used to define the orientation of the molecule.

In Figure 1.9 the structure of the bases in DNA is shown with the two types purines and pyrimidines illustrated. Purines have two rings, whereas pyrimidines have one ring. Bonds are represented by straight lines. In the bases, a carbon is present (and not shown) where two lines intersect. Bonds with no atom at the end have a hydrogen atom at the end and are hydrogen atoms.

A single strand of the DNA molecule is formed by the sequence phosphate-sugar-phosphate-· · · ·-sugar, with the $5'$ carbon of the sugar linked to the phosphate and the $1'$ carbon linked to the base. See Figure 1.10. Note that there is a definite direction to the chain, conventionally noted as $5'$ to $3'$.

Two chains joined by hydrogen bonds between so-called complementary bases form the DNA molecule. The complementary bases form the basepairs A-T and C-G in Figure 1.11. The hydrogen bonds are shown by dotted lines.

Finally we are ready to join the two strands into a complete DNA molecule in Figure 1.12. The strands must be of opposite orientation and, for a perfect fit, must be of complementary sequence. Given a sequence of one strand, it is obvious how to predict the other strand.

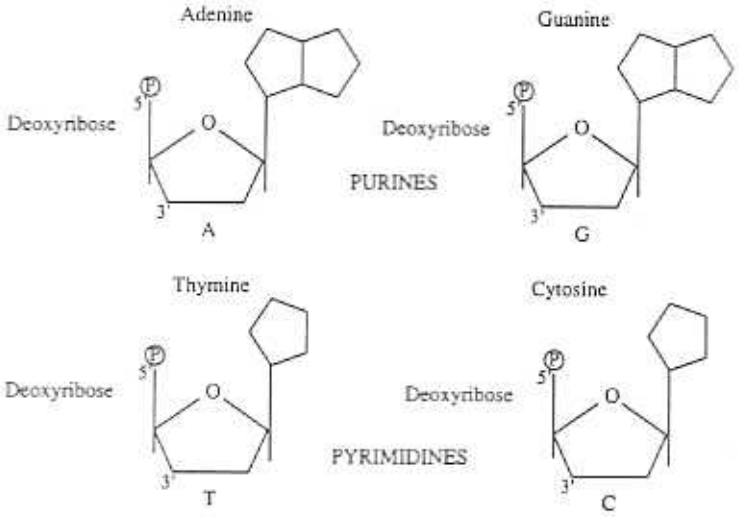


Figure 1.9: *Purines and pyrimidines*

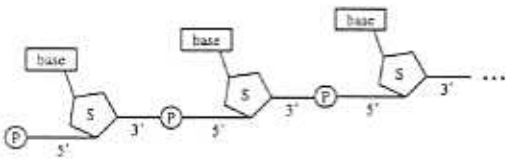


Figure 1.10: *DNA molecule*

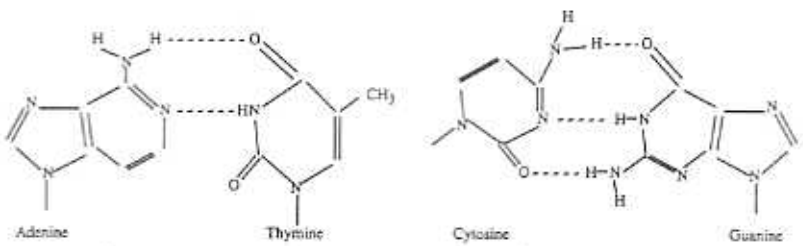


Figure 1.11: *Basepairs*

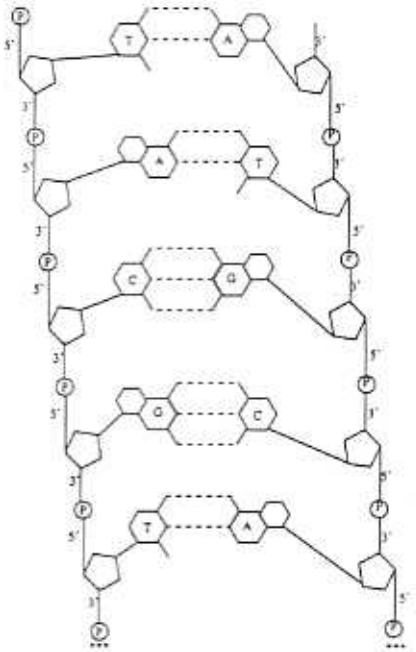
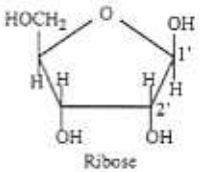


Figure 1.12: *Complete DNA molecule*

RNA

To describe ribonucleic acid (RNA), there is not much to do formally. The sugar in RNA is ribose instead of 2-deoxyribose.



Ribose

In place of thymine, we have the pyrimidine base uracil.