9

Molecular Genetics—the Chemical Basis of Heredity

In keeping with the theme of the book, enough genetics will be introduced to allow us to follow the role of energy in the propagation of individuals and species. We will still be speaking in general terms in this chapter, rather than discussing specific groups of animals or plants.

The earliest forms of life on Earth were single-celled, possessed no discrete nucleus and as such are called prokaryotes. Single-celled organisms which do possess a membrane-enclosed nucleus, the eukaryotes, developed over time from these simpler life forms. Eventually, evolution led to complex organisms such as mammals and flowering plants. These consist of a number of organs, such as the heart, liver, brain, skin, bark, leaves, and roots, which in turn consist of specialized groups of cells called tissues. The individual cells making up the various tissues and organs of these organisms can be considered the basic units, or modules, of complex life. There were advantages to becoming multicellular, but there were also additional problems to be solved. Fortunately, there was built into the primordial life forms the adaptability that was to allow them to evolve into more complex organisms.

Chromosomes were first observed in 1882 by Walther Flemming as tiny threads in the nuclei of dividing salamander larvae cells. It was established that the number of chromosomes in each cell of a particular organism depends on the species to which the organism belongs. Genes, the functional units that contain the code to make the entire organism, are arranged in linear fashion along the chromosomes. We now know that genes consist of molecules of DNA, which typically in eukaryotes are entwined in a framework of proteins that assists them to fold up into extremely compact structures. All of the somatic cells, or body cells, in animals and plants contain the full complement of chromosomes and hence the full complement of genes. In sexually reproducing organisms, the sex cells, or gametes, usually contain half the number of chromosomes found in somatic cells. When sexual reproduction occurs, half the chromosomes of the resulting embryo come from the sperm of the male parent and the other half from the ovum of the female parent, so that the embryo inherits characteristics of both parents and has a complete set of chromosomes.

Genetics is the study of inheritance and of genes. Details of the molecular nature of heredity have been revealed steadily over the last 200 years or so, and most aspects are now known in detail at the level of molecules. The science of genetics was well developed long before genes were actually proven to exist.

The Austrian monk Gregor Mendel (1822–1884) carried out breeding experiments with plants. Mendel was by training a physicist, encouraged by his abbot of the Augustinian monastery in Brunn (now Brno) in Moravia (located in the current Czech Republic) to study methods of agricultural improvement. In his pioneering work with sweet pea plants, he showed that inheritance takes place by means of discrete particles, which Mendel termed 'factors' and were later called genes. Thus, different factors were shown to be responsible for certain physical characteristics of the sweet pea, such as flower colour or seed shape. Previously, it was believed that characteristics in the offspring were inherited by a kind of 'blending' of the characteristics of each parent, by some unknown mechanism. Mendel showed that inheritance involved discrete units, the genes, which themselves remained intact. One of his greatest insights was that genes occur in pairs and that one member of each pair is inherited from each parent.

Mendel used his skills in mathematics to produce a theory of inheritance, first reported in 1865, which explained his quantitative plant-breeding results. It was largely ignored by biologists. At the time, his 'factors' were abstractions, the proofs mathematical, and the journal where he published the results obscure. Mendel's work lay dormant until 1900, when it was 'rediscovered', independently, by deVries, von Seysenegg, and Correns. At the time of Mendel, and indeed until 1953, the actual nature of genes was unknown. The very existence of genes as real chemical entities was debated for many years. For some time it was believed that genes were in fact special proteins. In 1902, Walter Sutton was first to point out the connection between chromosomes and Mendel's factors, but it was not until the 1930s that genes were shown conclusively to reside on the chromosomes in the cell nucleus, after important advances coming from studies in the USA of fruit-fly genetics by Thomas Hunt Morgan and later by Barbara Mc Lintock on maize chromosomes. Studies up to about this time are generally known as classical genetics. It was believed that the structure of DNA was too 'simple' to allow the encoding of the enormous amount of information for a whole organism. This belief proved to be wrong. When James Watson and Francis Crick published the structure of DNA in 1953, the science of molecular genetics became possible. The actual molecules that held all the information for life had been discovered. In a frenzy of scientific activity, it was soon shown that DNA possessed all the requirements for replication (i.e. to reproduce itself) and was also coded to carry all the information to produce an entire organism. The cells of all organisms, from bacteria to humans, carry DNA unique to the species. This fundamental complement of DNA carried by the members of each species is called the genome of the species. The genome of many species is divided into chromosomes, each of which consists of DNA plus various other structural bits and pieces, in mammals these being mainly special proteins called histones. Humans possess 46 chromosomes, in two equivalent sets of 23, in each somatic or body cell. The sperm and ova, that is the sex cells or gametes of humans, each contain only 23 chromosomes, so that when fertilization occurs, the full complement of 46 chromosomes in the form of 23 pairs, half from each parent, becomes part of each cell in the developing embryo.

Each chromosome is a long, continuous DNA molecule. A chromosome contains thousands of special functional regions, the elusive genes, and also

DNA regions of uncertain function. A number of scientists proposed that some of these regions actually have no useful function and called such regions 'junk DNA'. More recent opinion is that this dismissive attitude was probably premature.

Mendel was correct. Each gene has a specific function, or more than one function in some cases, that helps to determine characteristics such as flower colour, eye colour, sex, skin colour, etc. Some characteristics are controlled by more than one gene, so we can't say 'there is *a* gene for skin colour' because there is more than one gene controlling this characteristic. Similarly, it is probably premature to say there is a single defective gene responsible for asthma or for schizophrenia or for alcoholism. The full information is as yet unavailable. It is also likely that many diseases are partly of genetic and partly of other (environmental) origin, so the so-called 'gene therapy' envisaged by some genetic engineers is unlikely to be able to cure all our ailments.

The genome, the total complement of genes working together, is able to determine all the characteristics of the organism that contains it. The human genome contains some 20,000–25,000 genes, while the colon bacillus *Escherichia coli* has about 4000. In general, the more complex the organism, the larger is its number of genes (as one would expect intuitively) because there needs to be more coding to define fully the more complex structures and functions. Paradoxically, however, the total amount of DNA in a cell varies enormously between species. Humans have about 3.3 billion base pairs (see later) of DNA, while the single-celled *Amoeba dubia* has a staggering 670 billion base pairs. As noted above, not all of the DNA of either species is in the form of active genes.

Not all the genes in a particular tissue type are capable of being activated or expressed, to use a genetic term. The particular group of genes that is capable of being expressed in liver cells differs from the particular group of genes that may be expressed in brain tissue cells, even though each type of cell contains the full complement of genes. In fact, the pattern of gene activity in a tissue determines the type of tissue it is, or vice versa depending on how you like to think of it. This is one of the directions in which complex organisms have evolved and we can see why such a development is useful. The evolution of cells capable of different gene expressions, and of course the appropriate mechanisms for controlling such expressions, was essential to the development of specialized tissues and organs. The cells of a developing embryo undergo a process called differentiation, during which the previously undifferentiated cells (essentially, these are cells that are 'uncommitted' at the early stage of development) become specialized into groups, which on further development form into tissues. Eventually, in the fully developed organism, each specialized organ, such as the brain, the liver, or the heart, consists of its own types of tissue, made up of one or more types of differentiated cell.

Just to complicate things further, in any given mature tissue, not all the genes capable of being expressed are actually expressed all the time. The expression of all the genes, all the time, in fact would be disastrous, as not all the chemicals produced by genes are needed all the time. Strict control of gene activity is essential, as energy and many biological molecules are too precious to waste and are usually difficult to store. Genes are usually activated by feedback

mechanisms (see Chapter 11) so if the level of a particular gene product (such as an enzyme) should fall below the required level, this will be detected and the gene will be 'switched on', 'activated', or expressed, for example by the action of a hormone or a key metabolite. When the enzyme reaches the required level, the feedback mechanism will ensure that the gene activity will maintain it as long as required (see below). The mechanisms for control of expression of some microbial genes have been studied intensively and the general principles are also known for more complex organisms.

What is the most obvious common feature of all successful species? They and their ancestors have developed effective means of reproduction and survival. New species evolve from pre-existing ones and so on back through time. This constitutes the basis of the theory of common descent. A gene (DNA) from one species can be inserted chemically into the DNA of another species by so-called genetic engineering. Not only that, but if the insertion is done with the appropriate precautions, the gene can be induced to form its specific protein product (s), demonstrating that the chemistry of replication is similar for all life on Earth. Replication refers to the ability of DNA to make exact copies of itself. By implication, it also means that DNA-based replication is flexible, in that it succeeds for all the enormous variety of species which inhabit this planet. It succeeds for organisms which reproduce asexually or sexually, for plants and mammals, fish and fungi, for bacteria and beetles, and has done so for billions of years. As far as we know, DNA and its structurally close partner RNA are the only molecules generally used by living organisms for the process of replication. Evolution on the whole tends to be conservative, maintaining the systems that work well, rather than having to invent new ones.

The replication of DNA is only a part of the process of reproduction, which in a mammal involves fertilization of the ovum by a sperm, implantation in the uterus, development of the foetus, birth, and development to maturity. The fascination scientists have with DNA lies in the fact that, as far as information is concerned, everything is there. Within a single, identifiable structure, wholly accessible to modern science, lies the entire code for each finished organism, and apparently this code can be fully interpreted. This is the biologists' Rosetta stone, characterized by Watson and Crick a mere five decades ago and decoded soon after by others. Not only that, but all living organisms use essentially the same code. No wonder that the Human Genome Project, with its predicted completion date of about 2003 blown away by 2001, continues to cause excitement in both the scientific and the general population.

Modern genetics has provided a unifying principle that links all parts of biology. Previously disparate fields of biology can now be related by the study of DNA and genes. Areas such as physiology, developmental biology, morphology, ecology, and phylogenetic studies have all been advanced by an understanding of the underlying genetics. The term molecular evolution has recently come into existence and this field of study uses gene sequences to document the history of life on Earth.

Before I proceed further, let me emphasize that DNA isn't everything. By this I mean that although DNA carries all the information that eventually will produce an entire organism, this is not the end of the story. The development

of an adult mammal from a fertilized egg is an extremely complex process, which is still far from being fully understood. What happens at each of the many steps on the way to adulthood is not necessarily under the exclusive control of the genetic material. There is growing evidence that some of the developmental processes are dictated by what has been called 'necessity'. I sometimes describe evolution as the art of the possible, meaning that evolution by natural selection (or indeed by any other means) could only act, in a constrained way, on what was already there. Here I am saying something similar about events at the molecular level. At any stage of development, what an embryo can 'do' next is limited by its physical and chemical surroundings. The DNA in the genes cannot 'tell' each cell in the embryo *exactly* what to do, even though it may communicate via numerous chemical messengers to various cells. To some extent (as yet we don't know to what extent, or to what extent it may vary between species) the DNA loses exclusive influence, and to some extent the environment takes over. This realization has allowed biologists to look more broadly at development in an effort to solve hitherto intractable problems.

The environment in this case comprises the immediate physical and chemical surroundings of the developing cell or foetus, which of course are continually changing during the developmental process. The surroundings of a cell constantly send chemical messages, some of which directly or indirectly influence particular genes either to become expressed, and begin manufacturing their product, or to 'switch off' and cease to produce their product. This is an example of feedback control, which, as well as affecting genes, is common in most metabolic pathways (see Chapter 11). Maintenance of the correct ensemble of chemicals, at appropriate concentrations, around a developing organism is essential. The composition and concentrations of such compounds vary with the stage of development. In contrast, an imbalance, or the presence of unwanted chemicals in the surroundings of a human foetus, may have drastic effects on its development. A woman unfortunate enough to contract the viral infection rubella (German measles) during the first 8 weeks of pregnancy may have a child who has hearing defects, heart abnormalities, or liver disorders. It has been shown that some infections, rubella being one, some kinds of drug taking, and cigarette smoking during pregnancy can cause defects in the development of the foetus with consequences that may be life-long.

The extent and type of problem which may arise depends on the stage of pregnancy at which the foreign influences occur. Different organs develop at different times during a pregnancy—the two sides of the brain develop at slightly different times, as do other parts of the nervous system, the limbs, etc. Which organ is affected depends on the timing of an infection or on the ingestion of some foreign chemical compound that can penetrate the placenta and reach the foetus. There is medical evidence suggesting that infection in the mother during a period when the development of the two sides of the brain in the foetus is out of phase (i.e. one side is slightly ahead in development compared with the other) can result in an imbalance of receptors for the neurotransmitter dopamine. The left and right brain hemispheres may develop a difference of up to 20% in the number of dopamine receptors. This imbalance,

according to some medical opinion, may be a major contributor to the condition of schizophrenia, although this interpretation is by no means universally accepted by researchers.

A point related to what I have been saying about development is that:

The genome is more like a recipe than a blueprint or plan.

A blueprint or plan will show exactly the way the object will look in its final from. All detail is shown in the directions and the diagrams. In the case of a house, so long as standard building procedures are followed, its shape and appearance will be as shown on the plan.

In contrast, the shape and appearance of an animal or plant (its phenotype in biological terminology) is not obvious from its genetic makeup (its genotype) alone. The full sequence of all the DNA (i.e. the entire genome) of some organisms has already been determined. The first free-living organism to have its genome characterized was the bacterium *Haemophilus influenzae*. Another example is the bacterium *E. coli*, which is found in the intestines of us all. All their genes have been mapped, that is their locus positions on the DNA are precisely known and the entire DNA base sequence has been determined.

At the beginning of the year 2000, over 20 prokaryotic and some three eukaryotic genomes had been published, with the sequences of hundreds of more species expected over the next few years (*Science* 287, (2000) 605–6). Until very recently there was by no means a consensus on the number of human genes. Experts differed in their opinions and estimates ranged from as low as about 30,000 to nearly 200,000 genes in the human genome. Celera, a US company, and the multinational study called the Human Genome Project arranged to have a preliminary human genome sequence announced simultaneously in June 2000. In February 2001, the groups simultaneously published in the journals *Nature* and *Science* more details, including the estimate of 32,000 genes for the human genome. This was at the lower end of the previous estimates and one of the implications is that there is probably not a specific gene for every human characteristic. Rather, there are probably more control genes, which fine-tune the timing and extent of expression of other genes, than previously thought. Current reports suggest an even lower number of human genes, perhaps 20,000–25,000.

Another proposal is that the DNA of the same gene may be processed to yield more than one messenger RNA, and hence more than one protein. Despite much detailed sequence knowledge, it is not as yet possible to predict the phenotype (shape) of an organism from its DNA sequence alone. The analogy of a recipe for the genome now becomes clear. A recipe will not tell us the shape of the final cake. This will be determined by the shape of the baking dish (roughly analogous to the species to which the organism belongs). The recipe will hopefully tell us the approximate taste type and even give some idea of the texture of the cake, so long as the cook is competent. If there is an unexpected change in the environment during cooking (analogous to the change in environment during foetal development) then the final product may not turn out as expected in these respects, for example opening the oven door or using too high or too low a temperature can have drastic effects on the final taste and texture

of the cake. To push the imagery about as far as possible, we could say that a genetic (DNA) defect in a foetus is the analogy of an inaccurately followed recipe, leading to, say, cystic fibrosis or diabetes. On the other hand, a problem (e.g. rubella) occurring during foetal development resulting in a baby that has physical defects or behaves abnormally is analogous to a cooking error that results in a cake that looks or tastes unusual.

I won't go further into the problems of inherited diseases and developmental problems here. I want to describe the way in which the DNA in cells carries out two essential processes for the continuation and the sustenance of life. The first is replication of the genetic material (the genome) and the second is translation of small parts of the genetic message we have called the genes into something useful. Each type of gene is potentially capable of producing a specific type of protein chain, which may then do its designated job. I say potentially capable because as we have seen, each gene in each cell is not always active, that is it is not always expressed. A little thought will show that a careful balance between the thousands of possible gene products is essential. The maintenance of this balance is managed, in part, by controlling the 'on' and 'off' messages to the genes. In, say, a mammal, how does the body control this switching of genes? It is usually achieved by some kind of feedback mechanism, for example when the concentration of insulin (a small protein involved in glucose uptake by cells) in the blood becomes too low, this is detected by cellular sensors, which send a 'message' or 'signal', usually a chemical, to the special cells in the pancreas that produce insulin. The signal informs the cells that insulin is low and soon the cell activates (or, more accurately, stops the repression of) genes that control insulin synthesis and insulin is produced for transport to the bloodstream. Once the insulin in the bloodstream reaches normal levels, this too is sensed and the message goes all the way back to the pancreas cell genes, which are switched off till required again. You can imagine that this feedback loop is part of an incredibly important balancing mechanism that must be precisely controlled. There are many hundreds of other essential processes that need to be controlled by similar means. The internal balance of a complex organism such as a mammal must be monitored and maintained at all times. The interplay of molecules and energy is involved in all these processes. As we have seen previously, cells require a constant input of energy to maintain their physical integrity, to carry out metabolic processes, and thereby to remain alive. Energy is involved in the recognition processes taking place between the molecules involved in information transfer. The situation in single-celled organisms, which lack specialized organs and tissues, is less complex, but a high degree of control is necessary nevertheless.

How does all this work? What is so special about DNA? The full story is a fascinating one, which would take much space to describe, so I shall cut technicalities to the bone. DNA is a large molecule. Rather than referring to the size of DNA in Daltons, molecular biologists usually refer to the number of base pairs. The human genome has some 3.3 billion of these base pairs, so sequencing this enormous number correctly is a mammoth task. Each member of a base pair occurs on a different strand of DNA, the two strands being wound into the famous double helix. There are four bases in DNA: adenine (A), guanine (G), thymine (T), and cytosine (C).

An abbreviated way to represent the sequence of the bases on a single strand of DNA is

—AGTCAATGCTCAGGCTA—, etc.

We shall see later a similar way to represent double stranded DNA.

The sequence or order of bases in DNA in a particular gene is of extreme importance, as it determines the structure of the protein that the gene codes for. The base sequence is referred to as the primary structure of DNA. The maintenance of the correct sequence of bases is so important that each cell invests a considerable amount of its precious energy in DNA repair. Should an inadvertent change in the sequence in DNA occur, the structure of its protein 'offspring' could be changed, perhaps drastically. This is what happens in some inherited diseases such as cystic fibrosis, and in sickle-cell anaemia, which infects hundreds of thousands of people in parts of East Africa. Such inherited problems are often referred to as molecular diseases, and the precise causes of many are now known. The causes of many more molecular diseases will become known as the benefits of the Human Genome Project begin to flow through to medical science.

At the molecular level DNA consists of two intertwined double helices, with specific base-pairing holding the two strands together. What holds them in this highly organized, and apparently unlikely, conformation? The answer is hydrogen bonding between the base pairs. Figure 9.1 shows that the base pairs bridge across the two strands of DNA rather like steps on a spiral staircase. The double helix is referred to as the secondary structure of DNA (Figure 9.1).

The steps, however, have gaps in the middle of the tread that are occupied by hydrogen bonds; these are essential stabilizers of the double helix. Each base-pair step contributes the energy of either two (A-T) or three (G-C) hydrogen bonds to stabilize the double helix. We saw in Chapter 5 that many weak hydrogen bonds (about 20 kJ mol^{-1} in water) working simultaneously are able to stabilize a particular molecular conformation quite successfully and that is what is happening here. One of the great insights of Watson and Crick was to make their models of the four bases A,T, G, and C in the correct form, such that the hydrogen-bonding possibilities became obvious. The insight came, after much unsuccessful and frustrating model building, after one of them had spoken to a chemist who understood that some molecules are capable of existing in two forms, called tautomers. It so happened that for most of their model building Watson and Crick were using the wrong tautomeric form for the bases. As soon as the correct forms were used, the solution essentially 'fell out'. They saw that A and T could form two hydrogen bonds, while G and C readily formed three (Figure 9.1). However, A didn't readily form such bonds with G or C and neither could T.

The specific pairing of A-T and G-C also explained an experimental finding that had puzzled biochemists for many years. It was well known that in DNA, the molar amount of A always equalled that of T, and similarly the molar amount of G equalled that of C. The specific pairing noted above shows why. The Watson/Crick double helix in fact explained a number of things, and several of these involved specific hydrogen bonding.

Firstly, it explains replication, which is the process during which DNA makes copies of itself and which makes reproduction of the whole organism possible.

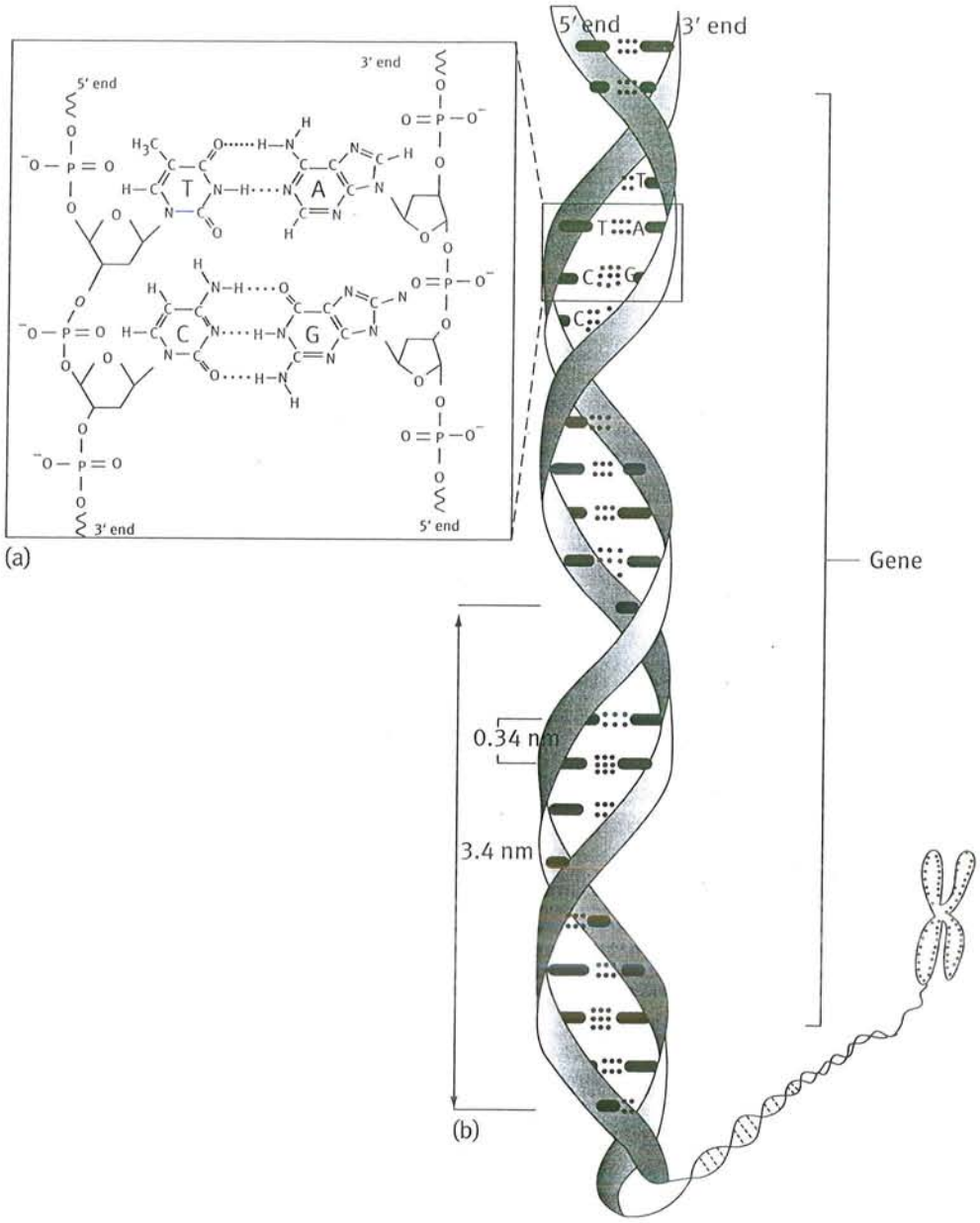


Figure 9.1. (a) Base pairing in DNA. (b) From double helix to metaphase chromosome.

Offspring need their own copy of the genome and this comes about by means of DNA replication. The process involves base pairing directed by hydrogen bonding, illustrating the highly specific role of energy at the most fundamental level of life.

The average length of DNA in a human chromosome is about 50 mm. The entire genome is the equivalent of about a metre of DNA. Each somatic cell in your body (a typical cell being about 200 microns (1 micron = 0.000001 or 10^{-6} m) across) contains the full genome. All the DNA must fit into the cell nucleus, which at most is only about 5 microns (0.000005 m) in diameter. This is a most amazing packaging feat. In the broadest sense we can readily believe that the orderly packing of DNA into the nucleus is a process that involves a decrease in

entropy, and this overall will be energy consuming. The supercoiling of the DNA double helix is known as the tertiary structure of DNA.

Replication of DNA is shown in Figure 9.2. There are many details to this process, all of which are known essentially in full, but we will concentrate mainly on the energetics of the major steps. DNA replication involves separation of the two strands and, using these as templates, the production of two new strands. As half of the original DNA molecule is conserved in each replicated molecule, this process is called semi-conservative replication, which takes place

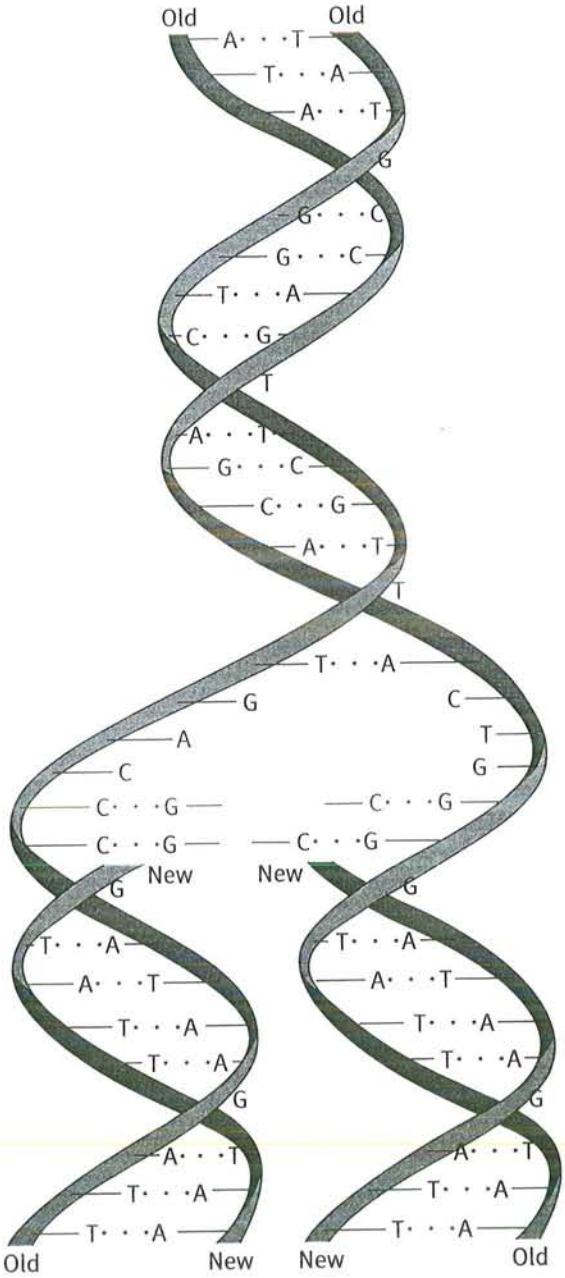


Figure 9.2. Semi-conservative replication of DNA. (Illustration from Voet, D., Voet, J.G., and Pratt, C.W. *Fundamentals of Biochemistry*, upgrade edn, p. 54. © 2002 John Wiley & Sons Inc.)

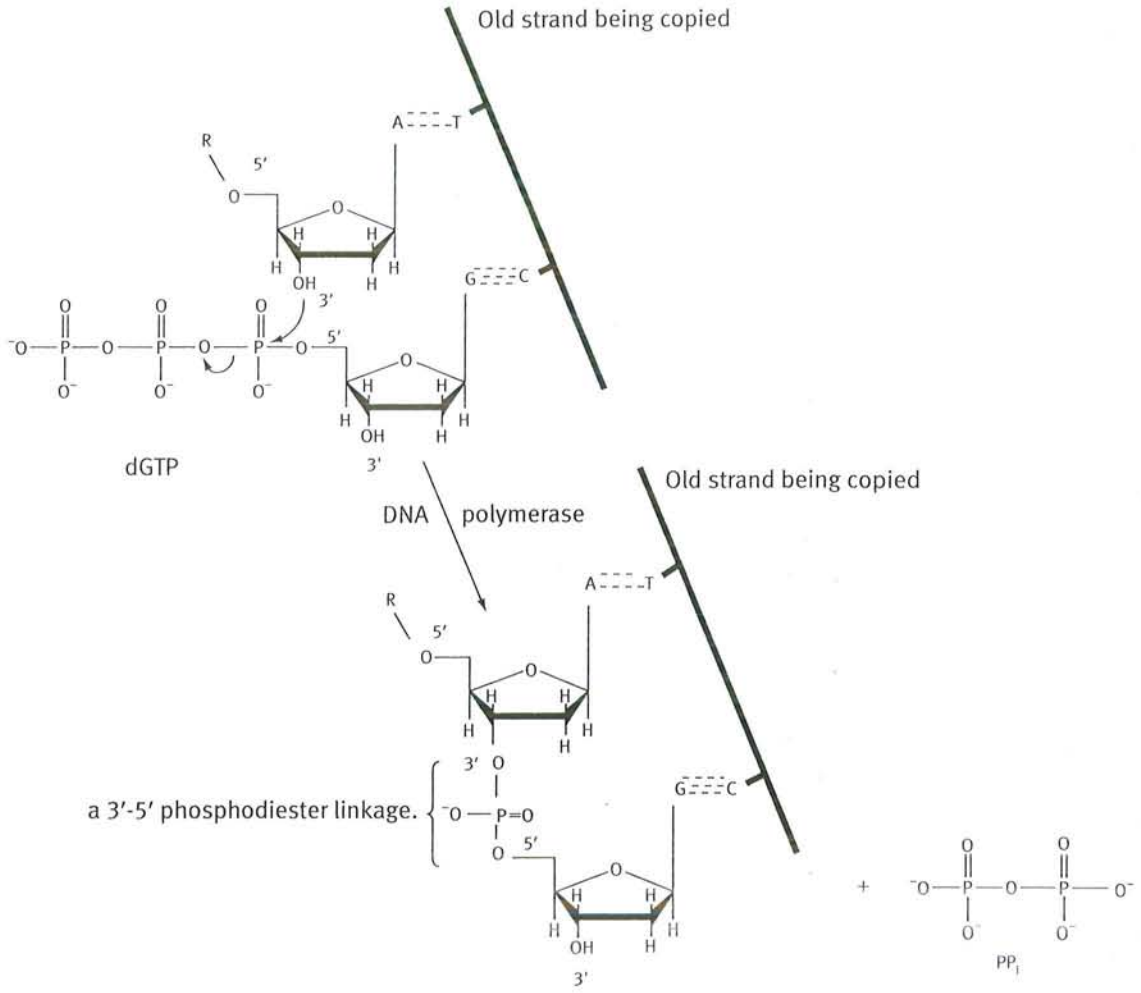
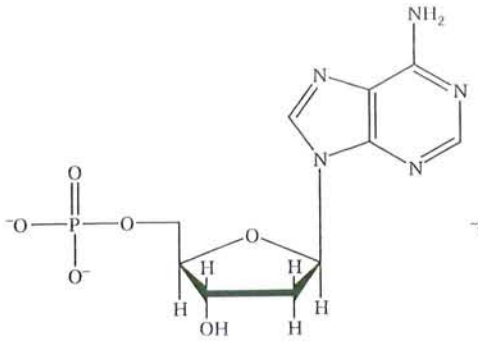


Figure 9.3. Replication of DNA. Extension of the new DNA strand by formation of a 3'–5' phosphodiester linkage. The reaction is driven by the ΔG of hydrolysis of dGTP, with the release of pyrophosphate (PPi).

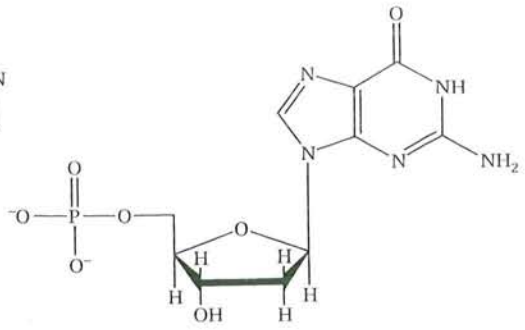
in both prokaryotes and eukaryotes. The description is a general one, illustrating the principal steps common to most organisms.

The enzyme performing the actual replication is called a DNA polymerase, of which there are up to five types in prokaryotes, for example *E. coli*, and five (slightly different) types in eukaryotes. Overall there are several other enzyme types involved in replication. A DNA gyrase introduces a swivel point ahead of the replicating strands and a helix-destabling protein (a helicase) binds and promotes unwinding. The single-stranded regions of the unwound DNA are stabilized by a single-stranded DNA-binding protein, which also protects the single strand from hydrolysis. The helicases possess ATPase activity, which harnesses the energy of ATP to break the hydrogen bonding between base pairs and separate the DNA strands. Two molecules of ATP are consumed for each base pair separated, so full replication requires substantial amounts of energy.

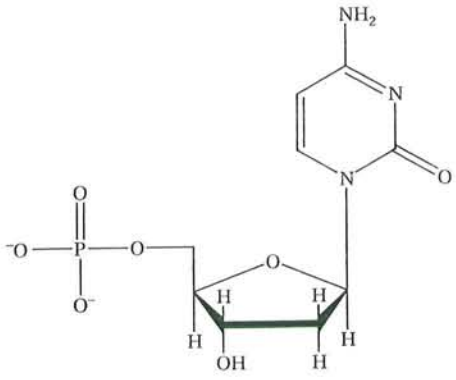
Further energy is now involved in the formation of 3'–5' phosphodiester linkages to extend the new strand. One of the four trinucleosides dATP, dTTP, dGTP, or dCTP will be added. For example, the G of dGTP hydrolysis is harnessed via a polymerase enzyme to couple it to the growing new strand. (Fig. 9.3). This process is continued until he entire piece of DNA undergoing replication has been completed.



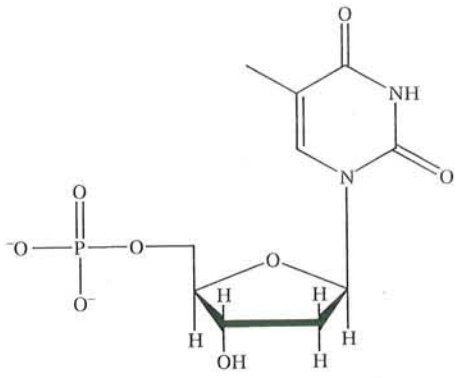
dAMP



dGMP

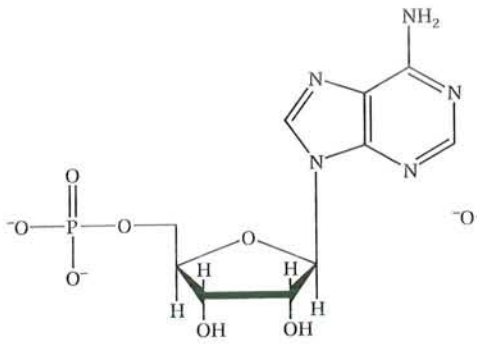


dCMP

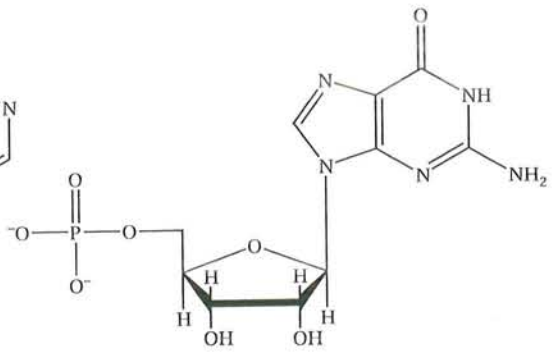


dTMP

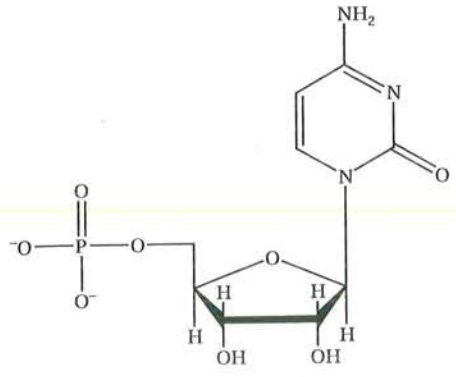
The prefix 'd' stands for 'deoxy', to distinguish them from the ribonucleoside phosphates, such as AMP, UMP, GMP, and CMP:



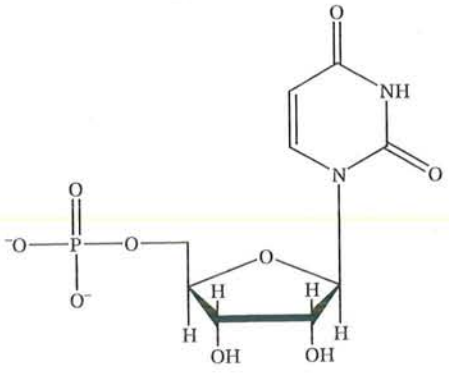
AMP



GMP



CMP



UMP

(The ribonucleosides, the precursors of the other major type of nucleic acid, RNA, differ from their deoxy equivalents only in the type of sugar they contain: ribose instead of the very similar, but importantly different, 2-deoxyribose. Like DNA, RNA contains four bases, the familiar A, G, and C, but instead of T it has U (uracil). Apart from its sugar group, RNA is structurally like a single-stranded DNA). All the deoxyribo- and ribonucleoside triphosphates are 'high-energy' compounds like ATP, that is they have a large negative standard Gibbs energy of hydrolysis, ΔG^0.

Although replication of DNA is essential for propagation of the species, for an individual member of that species it is essential to access the DNA that has all the information needed for its life, but this coded information must be decoded and processed into whatever molecules are required and precisely when they are required. When a gene is expressed, the part that codes for the product (usually a protein) must be copied to form a corresponding RNA. This RNA, called messenger RNA (mRNA) is then moved outside the nucleus to special structures called ribosomes, where it directs the actual protein synthesis. For the DNA of a gene to be copied to form mRNA, other molecules must be able to have access to it. As was the case for replication, the double helix must be unwound at least temporarily so that the machinery for its transcription into mRNA can act. The 'message' is transcribed from one strand of the DNA only, the so-called template strand. The enzyme that unwinds the DNA double helix, catalyses the polymerization of new bases into mRNA, and rewinds the DNA is called RNA polymerase, of which there are three types in eukaryotes and one type in prokaryotes.

The mRNA acts as a messenger molecule, but other types of RNA have different roles, such as major structural elements of the ribosomes. We will encounter some other roles of RNA later.

During transcription, U on the growing RNA strand pairs by specific hydrogen bonding, with A on the template DNA strand, and G pairs with C. As the RNA is produced we can represent it in our familiar abbreviated form as AUGCCAUGCAU, etc., as shown in Figure 9.6.

There are special signals built into the transcription process to start and stop the synthesis of mRNA so that each molecule is exactly of the length and base sequence dictated by the DNA of its gene. The completed mRNA is transported out of the nucleus to structures in the cytoplasm of the cell, the ribosomes, which are the sites of protein synthesis. In the ribosomes, the mRNA itself is used as a template, directing amino acids (in an activated form, so there is energy required here also) to be joined in the precise sequence that was coded by the DNA back in the nucleus. The DNA has thereby 'masterminded' the synthesis of a protein with the precise sequence that is most suitable for its function in the life of the cell or organism. No wonder DNA must be protected from mistakes such as replication errors and/or mutations, which alter the base sequence. An altered base sequence means that there is the possibility (but not the certainty) that one or more different amino acids could be coded for and inserted into the protein, which could

alter the function of the protein. Examples where such mutations have occurred are the genetic diseases cystic fibrosis and sickle-cell anaemia, as mentioned above.

We have seen that expression of a gene leads via transcription from DNA to the corresponding mRNA, but precisely *how* does the mRNA become translated into proteins in the ribosomes?

This is the role of the now-famous genetic code. The genetic code is what is known as a triplet code, meaning that three bases on mRNA form the code for each amino acid.

There are 20 amino acids used to form proteins. Can four bases, A, U, G, and C, taken three at a time, form enough unique triplet codes for 20 different amino acids? Easily. The first base of a triplet can be any one of the four. The same is true for the second and third places of the triplet, leading to $4 \times 4 \times 4 = 64$ possible triplets, more than enough. What about the extra possible combinations that are 'not needed?' All amino acids have more than one triplet code. Some have up to six, and because of this occurrence of several triplets for the same amino acid, the genetic code is said to be degenerate. The coding triplets in the mRNA are called codons.

In many cases the degenerate codons for an amino acid differ in the third base, but this is not always so. Of the triplets, 61 code for amino acids. The one coding for methionine, AUG, also forms part of the 'start' or initiating signal for the first amino acid in the new protein chain. There are also three special triplets that code to 'stop' the protein chain (UAA, UAG, and UGA; Figure 9.4).

Note that while there are several triplets for each amino acid, there is not any triplet which codes for more than one amino acid. The reason is obvious. If this were the case, the ribosome would be 'in confusion' as to which of the two amino acids to add. Put another way, as the sequence of amino acids in a protein strongly affects its function, the chance that two amino acids could yield the same functionality is low. Just how low is shown by the fact that such a situation simply does not occur in living organisms. Other points to note are that the code is non-overlapping (no bases are shared between consecutive codons) and is essentially universal, that is it is the same for nearly all organisms, underlining the common origins of life on Earth. (For a long time the code was believed to be truly universal, but recently it has been shown that the mitochondrial genomes of some species use slightly different codes. Note that mitochondria have some of their own genes, that is these genes do not occur in the cell nucleus, but in the mitochondria themselves. The reasons for this will be discussed in Chapter 12.)

When mRNA binds to a ribosome, it is arranged in such a way as to have its coding triplets accessible to another type of RNA called transfer RNA, or tRNA.

In each cell there exists a 'pool' of tRNA molecules available for protein synthesis. There is one specific type of tRNA available for each amino acid. Thus, the tRNA for phenylalanine will become linked only to phenylalanine by a specific synthetase enzyme, and become linked in such a way that phenylalanine is in an

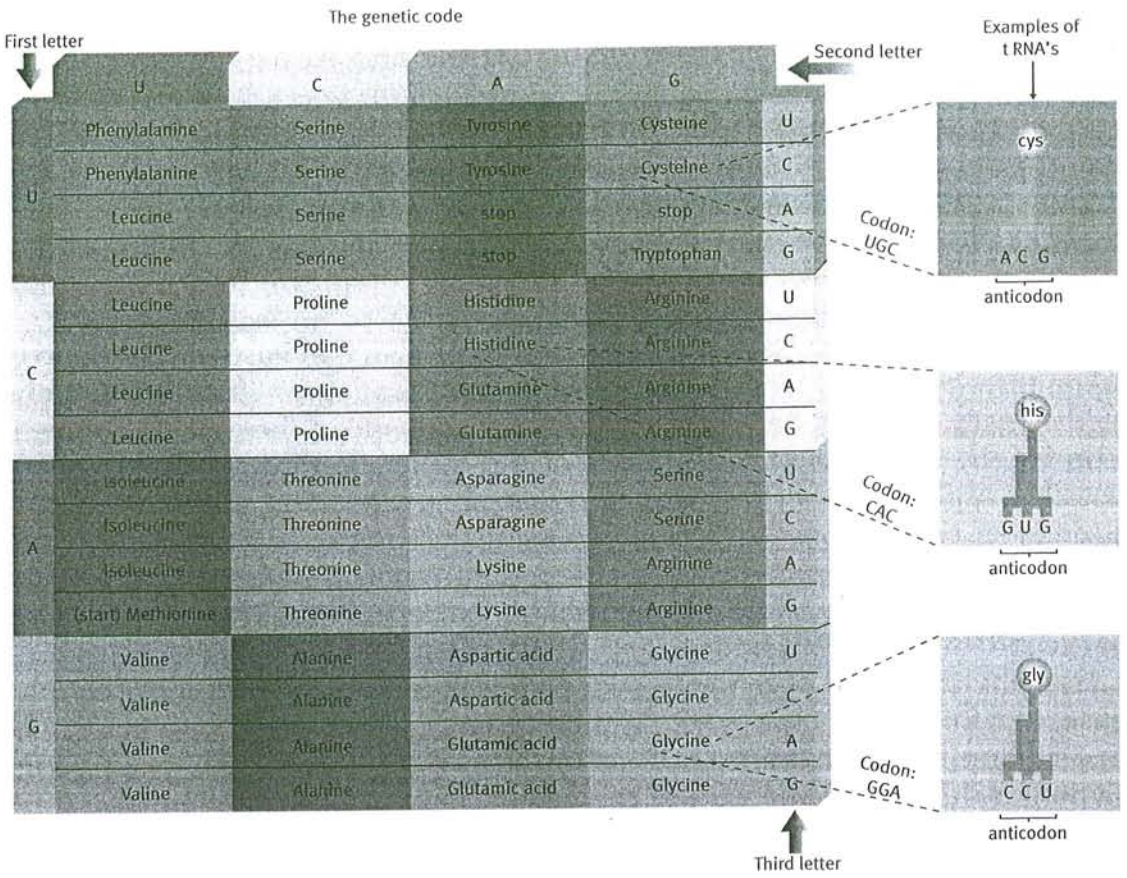


Figure 9.4. The genetic code. (Illustration from Trefil, J. and Hazen, R.M. (2001) *The Sciences—an Integrated Approach*. p. 520. © 2001 John Wiley & Sons, Inc.)

'activated' form, ready to form a covalent bond with the previous amino acid in the growing protein chain. This aminoacylation process consumes one mole of ATP per mole of amino acid (Figure 9.5).

The overall aminoacylation reaction is:

$$(\text{tRNA}^{\text{Phe}}) + \text{phenylalanine} + \text{ATP} \rightarrow \text{phenylalanine-}(\text{tRNA}^{\text{Phe}}) + \text{AMP} + \text{PP}_\text{i}$$

The details of the reaction are more complex than indicated. Note that the ATP forms AMP and PP_i (inorganic pyrophosphate) rather than the familiar ADP + P_i. Cells sometimes use this method of group transfer and the ΔG is similar.

How does the phenylalanine-tRNA 'know' its correct position on the mRNA waiting on the ribosome? This is where the genetic code performs one of its most important functions. On a region of the phenylalanine tRNA is a triplet sequence (AAG) that is complementary to the codon for phenylalanine on the mRNA (UUC). This triplet sequence is called an anticodon.

As a result of the specific pattern of hydrogen bonding between the UUC and AAG, the phenylalanine tRNA binds to the mRNA in the ribosome in exactly the correct position, the phenylalanine molecule is brought into close contact with the preceding amino acid in the growing protein chain (which is already bound

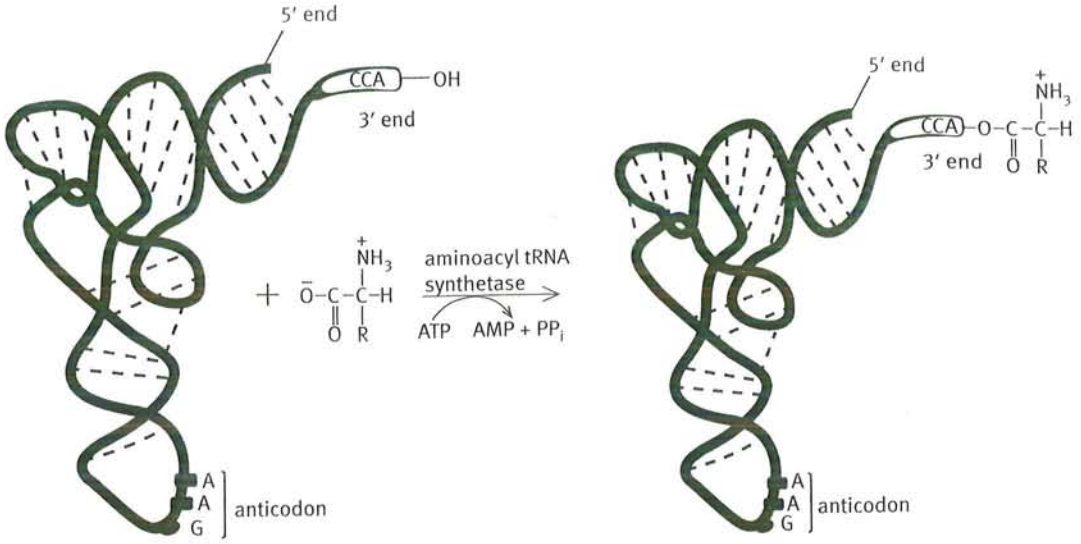


Figure 9.5. Structure of phenylalanine tRNA ($tRNA^{Phe}$) from *E .coli*, showing the tertiary folding and major hydrogen bonds (dashes). The anticodon for phenylalanine (AAG) is bottom left. Phenylalanine attaches to an adenosine at the 3′ end (top right) to form an aminoacyl tRNA, in readiness to add phenylalanine to the growing peptide chain being synthesized on the ribosome.

to the ribosome), and a covalent bond is formed, using the energy in the phenylalanine-($tRNA^{Phe}$) acyl bond to drive the process via a coupling mechanism. The ribosome then moves along the mRNA, which becomes ready for the next amino acid tRNA to bind, and the process is repeated. When the final amino acid has been added, the completed protein chain is removed from the ribosome and is ready for further processing. This further processing may involve cleavage of parts of the chain, as in enzyme precursors, the addition of special groups, such as haem groups in haemoglobin, folding into a specific three-dimensional shape as in many enzymes, or the addition of sugar chains (glycosylation), as in most of the blood proteins. The processing stages also require the input of energy via ATP.

A very simplified scheme of eukaryotic protein synthesis is shown in Figure 9.6. Note the direction of each DNA or RNA chain. The directionality, 5′ to 3′ or 3′ to 5′, is important. The expanded region shows mRNA having codons for phenylalanine, threonine, and serine. These are translated on the ribosome and are shown as part of the completed product of gene 2. Processing of the pre-mRNA to remove introns has not been included.

We have now seen the fundamentals of the way in which DNA is replicated, a process essential to the reproduction of organisms, and the way it is transcribed and finally translated into protein. It is worth reiterating the importance of the recognition processes between a coding strand and mRNA, and between codon and its anticodon. It is largely the energy of the hydrogen bonds between the appropriate base pairs that allows this to happen with such precision. The functions of DNA are essential, so it is protected from change or degradation as much as possible. Each organism spends considerable amounts of energy on

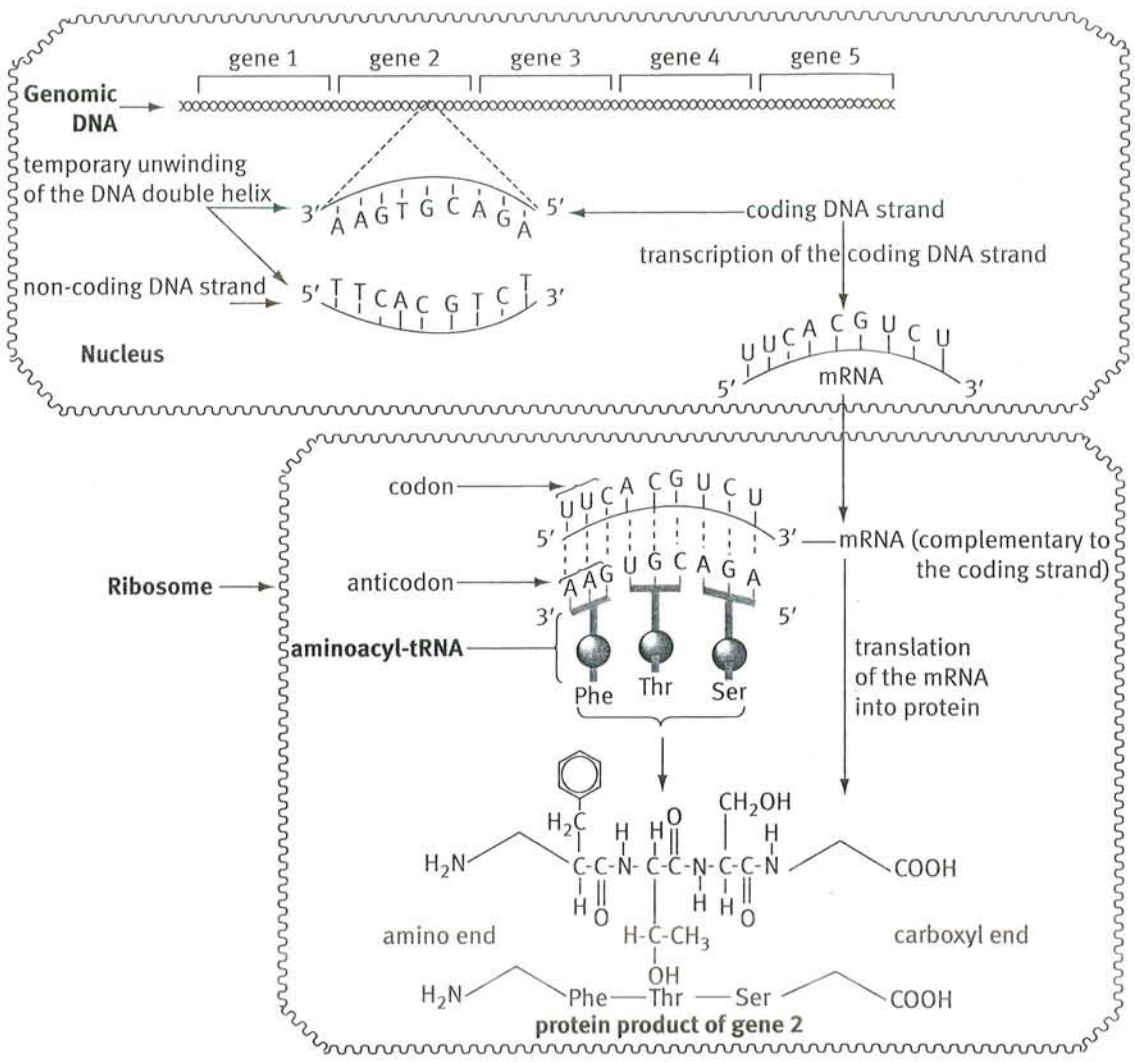


Figure 9.6. Protein synthesis in a eukaryotic cell. The nucleus provides a separate compartment for transcription. In eukaryotic cells, the original RNA transcript, called pre-mRNA, is processed in various ways before leaving the nucleus as mRNA. This is not shown. Translation occurs after the mRNA is transferred from the nucleus to the ribosomes.

DNA repair and on ensuring that the base sequence of its DNA is not altered. Such alterations lead to errors in replication, which of course affect the product of the gene to which the altered DNA belongs. Errors in hydrogen bonding that lead to the incorporation of the wrong base occur about once in every 10^4 to 10^5 base pairs and are called copying errors. In addition to copying errors, changes in the base sequence called mutations can occur. These may be caused by exposure to mutagens, agents which lead to mutations. Examples of mutagens include ultraviolet light, radioactivity (ionizing radiation), and certain chemicals. Other problems leading to genetic defects may be inherited. The copy of a 'defective' gene may be inherited from one or other parent, generation after generation. Well-documented examples include those of several European royal families, whose members suffered from an obvious disease

such as haemophilia. Ultimately, however, such defective genes must have arisen from mutation(s) in an ancestral 'normal' or 'wild-type' gene.

Fortunately, proofreading and repair mechanisms have developed over evolutionary time and come into action so that replication errors overall are kept to an absolute minimum, normally occurring about once in 10^9 to 10^{10} base pairs. Despite these very efficient protective mechanisms, changes in DNA have always occurred and will always occur. Defects in DNA repair processes can be disastrous. The disease xeroderma pigmentosum causes the development of skin cancers at an early age because the sufferers do not have the repair mechanism to correct damage to skin-cell DNA caused by ultraviolet light. An enzyme that cuts out the damaged portion of DNA is believed to be missing. The cancer eventually spreads through the body and usually causes death.

Mutations which end up in the gametes of one generation will, as long as they are not so serious as to cause immediate death or infertility in the offspring, appear in the next generation of a population. As a result, mutations are the basis of much genetic variation (diversity) in populations, which may or may not be advantageous. Genetic variation may act to protect populations from extinction. There are many examples of this, although not all are favourable in the eyes of humans. Bacteria and other microbial organisms develop resistance to drugs, insects eventually resist pesticides such as DDT, and unwanted plants avoid extinction by developing resistance to herbicides. This is not good news, unless you happen to hold shares in a drug or agrochemical company.

Crops that are too homogeneous genetically are highly susceptible to pathogens, such as potatoes in the nineteenth-century Irish famine and sugar cane strain Q124, which was widely used in the Mackay region of Queensland, Australia during the year 2000 growing season. Both crops were devastated. More genetic variation would have allowed increased numbers of pathogen-resistant plants to survive and might perhaps have averted the disastrous results. Mutations are also responsible for some cancers, such as skin cancers, which can be caused by high-energy ultraviolet radiation penetrating the nuclei of surface skin cells and altering the precious sequence of their DNA. Thus, as with many things in biology, there are several facets to genetics, at least when looked at from our anthropocentric viewpoint. Evolution is driven by mutations. As mentioned in a previous chapter, with no mutations there would be essentially no evolution of new species. The living world as we know it would not exist and, probably, neither would we.

Over evolutionary time, mutations have been recorded in the genomes of all species. One spinoff of the recent upsurge in the studies of genomics is its application to so-called molecular evolution. The work of Cambridge biologist George Nuttall and others in the early twentieth century eventually led to acceptance of the important principle that the degree of similarity between genes reflects the closeness of the evolutionary relationship between them. That is, if a gene sequence for one type of organism is very close to that of the same gene in another organism, then the two organisms are likely to be closely related. Comparative studies of gene sequences have been used to assist the more classical means of classification of organisms. Such studies, now termed molecular systematics, have led to the restructuring of the phylogenetic tree, which until the 1970s divided

cellular organisms into prokaryotes, which possess no discrete nucleus, and eukaryotes, which do. Karl Woese and his colleagues showed that there were in fact two distinct groups of prokaryotes, the Eubacteria (now called the Bacteria) and the Archaebacteria (now known as the Archaea). This work led to the recognition of three domains of cellular life—the Bacteria, the Archaea, and the Eukarya—and was based on sequence analysis of the 16S ribosomal RNA (rRNA). The three domains will be discussed further in Chapter 12.

The 16S rRNA, being essential to the synthesis of proteins in all organisms, is highly conserved. Being highly conserved means that it is so important a molecule that organisms with mutations in the 16S rRNA are unlikely to survive, with the result that very few mutations survive over evolutionary time in the gene that codes for this rRNA. This helped the researchers in that they had a simpler task in sorting out the implications of the relatively few differences in RNA sequence and led to the revelation of the two prokaryote groups quite early in the development of nucleic acid sequencing. The rate of DNA sequencing has developed greatly since the 1970s, with the complete genomes of several organisms from all three domains having been sequenced since 1995.

In contrast to that for the highly conserved 16S rRNA, other genes mutate at much higher rates. An example is that of the human immunodeficiency virus (HIV), which is the causal agent of AIDS. HIV evolves about a million times faster than the average human gene, which is one reason why the development of AIDS drugs and vaccines is such a problem.

A major feature of HIV evolution are the many and rapid changes in the carbohydrate/protein (glycoprotein) 'coat' of the virus, and these changes help HIV to stay ahead of developments designed to destroy it.

The long sequences of non-coding DNA that occur in many genomes (the badly named junk DNA) can 'tolerate' mutations to a much greater extent than can coding DNA. As the mutations cause no apparent harm, there will be little or no selective pressure to remove them and so such mutations accumulate in genomes, giving rise to a high degree of diversity in the junk regions. This diversity in the non-coding regions of the human genome is used in DNA fingerprinting or profiling to identify individuals, as the likelihood of two individuals having the same profile is extremely remote. The tests are often used for determination of paternity or for forensic purposes. As well as convicting a number of true criminals, DNA profiling has also led to the release of wrongly convicted persons, for whom the genetic triplet code and hydrogen bonding have come to the rescue. Having said that, recent legal challenges to DNA-based evidence have shown that the technique is not infallible.

Another area that is receiving benefits from gene sequencing was actually developed before DNA sequencing was technically feasible. This is based on protein sequencing. Historically, it happened that the techniques for the full sequencing of proteins were developed before those for full sequencing of nucleic acids. Incidentally, the same man was involved in both sequencing breakthroughs: Frederick Sanger from Cambridge University, who received two Nobel prizes, one for the development of each technique. As the amino acid sequence of a protein largely determines its three-dimensional structure and thus its function, it is therefore important to conserve the correct sequence.

The amino acid sequence of a protein is in turn determined by the messenger RNA that directed its synthesis, and as the mRNA sequence is determined by the DNA sequence of its gene, the protein sequence is maintained by the corresponding gene, as we have seen. Changes (mutations) in the gene DNA may be reflected in the amino acid sequence of the protein. As an example of all this, let's take the inherited disease sickle-cell anaemia. As we have seen earlier, haemoglobin is a protein in mammalian red blood cells that transports oxygen around the body. The way in which haemoglobin achieves this is known in great detail. Red blood cells are shaped rather like doughnuts, except that there is no central hole, and this shape is just perfect for the uptake and release of oxygen and several other simple, but important, chemicals, such as carbon dioxide and bicarbonate. Up to 20% of people in certain parts of East Africa have sickle-shaped red blood cells. These unfortunate individuals suffer from anaemia, as the misshapen red cells cause blocking of small blood vessels, thereby cutting off the circulation and causing tissue damage. The sickled cells are also more fragile and rupture easily, causing anaemia. What causes sickle-cell anaemia?

The answer lies in the amino acid sequence of sickle-cell haemoglobin (Hbs). It was shown experimentally that in position six of the two β-chains of haemoglobin (which is a tetramer of two α- and two β-chains) the normal glutamic acid residue had been replaced by a valine residue in the Hbs. Glutamic acid has a polar side-chain with acid properties, whilst valine has a non-polar hydrocarbon side-chain, which is quite hydrophobic in its properties. As a result, the Hbs crystallizes into long aggregates and in so doing pulls the red blood cell out of shape, causing the problem. This is another example of energy influencing molecular properties. Why did the Hbs crystallize? In general, and given the appropriate conditions, molecules may crystallize when their arrangement into crystals generates a lower energy state than the non-crystalline form. In this case, the deoxygenated form of the Hbs has hydrophobic pockets into which the valine side-chains of adjacent Hbs molecules pack nicely. The small interaction energy between the valine and the hydrophobic pockets lowers the overall energy level just enough to cause aggregation of a large number of Hbs molecules. The packing of Hbs molecules to form elongated crystals inside the red blood cell, and thus deform the cell wall, is an energetically favourable process, but it is unfortunate for the sufferers of sickle-cell anaemia. This is an example of the exquisite molecular detail in which it is now possible to study certain diseases.

What happened to change the glutamic acid to a valine in Hbs? If you look at the genetic code, you will see that the codons for glutamic acid are GAA and GAG, while two of those for valine are GUA and GUG. The substitution of U instead of A will change both the codons of glutamic acid to the two codons for valine, and apparently this is the mutation that occurred many years ago in the African population now afflicted with sickle-cell anaemia.

Molecular medicine is an apt term for studies on diseases such as these, and no doubt many more examples will emerge as a result of further work. At present more than 400 mutants of haemoglobin are known, most of which are single amino acid substitutions. Not all of these are harmful, and the availability to

scientists of such a variety of structures has been useful in structure/function studies of proteins. The haemoglobin molecule, being so widespread in living organisms, has also been extremely useful in molecular evolution studies.

Before leaving sickle-cell anaemia, let's consider a rather puzzling aspect. If sickle-cell anaemia is so dangerous to health, how is it that the sickle-cell gene persists in those east-African populations? On the basis of natural selection, one might expect the defective gene to be eliminated, as the survival rate of the parts of the population with the Hbs gene would be less than the survival rate of others in the general population without it. The reason proposed is that resistance to a virulent form of malaria appears to be closely linked to the Hbs gene. Resistance to malaria has been a survival advantage in Africa, so presumably its genetic component has persisted, dragging the Hbs gene along with it. On balance, the two features have allowed the survival, at least until now, of the deleterious gene.

I will use this example to illustrate some of the possible complications of the emerging gene therapy. Even though we will soon have the complete sequence of all the DNA in the human genome, plus the exact locations of all the genes, there will still be an enormous amount to learn. We will need to discover more about the control of gene expression and about the ways in which genes and gene products interact. Who would have anticipated the protective effect of the malaria-resistance gene on the Hbs gene, if indeed this is the full explanation? Scientists are now aware of this and other examples, and as the data comes in new solutions will be found for the problems that are bound to emerge. There is an exciting future in store for those involved in gene therapy.

As a final illustration of the role of energy in molecular genetics, I want to link one of the most ancient examples of life on Earth to one of our most recent scientific developments.

As part of the processes of genetic engineering and genetic fingerprinting, it is often necessary to increase the quantity of a particular section of DNA. A small quantity of the DNA may be obtained from some natural source or from the scene of a crime. The quantity may not be enough to enable all the necessary experiments to be carried out.

How can more of the identical DNA be obtained? In principle, there are three possibilities. Firstly, there is chemical synthesis. If the exact base sequence of the section of DNA required is known, there are automated techniques available to carry the synthesis out. They are relatively slow, complex, and expensive. Secondly, there is cloning, where the required piece of DNA is inserted into the genome of a suitable organism such as a yeast or a bacterium, the organism is grown in culture, and the DNA is isolated and purified. This is also a complex and expensive process.

In the 1980s Kary B. Mullis developed what is known as the polymerase chain reaction (PCR), which greatly improved on the first two procedures for the amplification of sections of DNA (Figure 9.7). For this he was awarded the Nobel Prize for Chemistry in 1993, together with Michael Smith, who developed the technique of site-directed mutagenesis.

In PCR a segment of double-stranded DNA containing the base sequence of interest is heated to about 95°C to separate the two strands. Remember that the

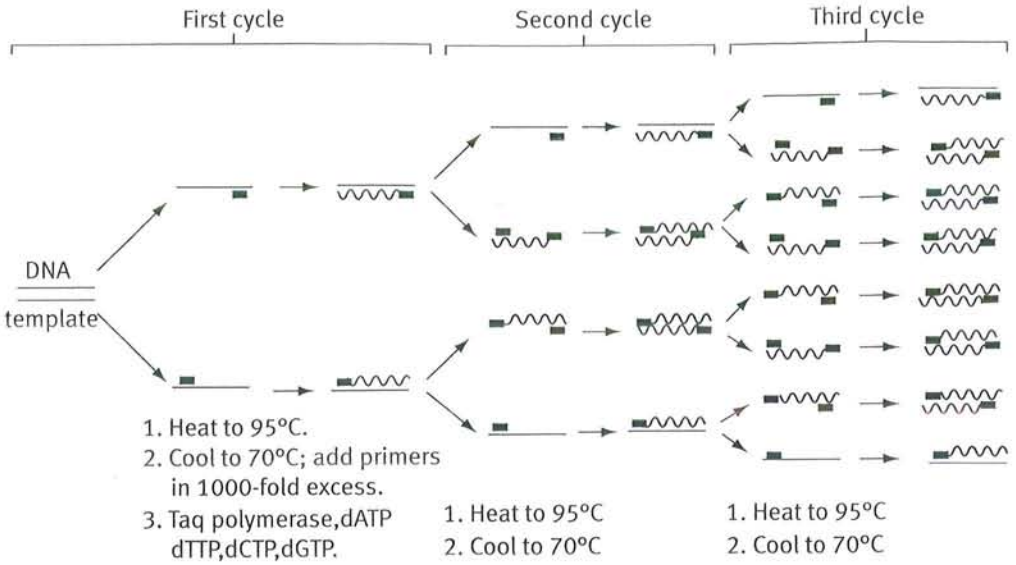


Figure 9.7. The polymerase chain reaction (PCR). Each cycle consists of three steps, as shown, though DNA, *Taq* polymerase, dATP, dTTP, dCTP, and dGTP are added only once. The small black rectangles represent DNA primer oligonucleotides.

two strands are held together by cooperative hydrogen bonds between A and T (two bonds) and G and C (three bonds). Heating to 95°C is sufficient to dissociate the hydrogen bonds without affecting the covalent bonds that maintain the primary structure of the DNA chain. The separation of the two strands is necessary because during PCR each strand is synthesized separately, in a kind of semi-conservative replication process similar to natural replication, except that both DNA strands are replicated simultaneously in PCR. A large excess (1000-fold) of short oligonucleotides, called primers, which have been chosen on the basis of their base-pair complementarity to the 3′ end of the DNA sequence chosen for amplification, is then added and allowed to cool with the DNA, a process called annealing. During annealing, the short oligonucleotide primers line up, their As to the DNA's Ts and their Gs to the DNA's Cs, etc. and so in their specific places adhere by hydrogen-bonding to each separate strand of the DNA. After annealing, therefore, we have each strand of the section of DNA destined for amplification bounded by these short primers, which in effect act as markers for DNA polymerase to start and stop DNA synthesis. DNA polymerases require primers to add to. They cannot synthesize DNA from single nucleotides such as ATP, etc. Our DNA template system for the synthesis of new DNA is now ready.

Also present is a suitable DNA polymerase, an enzyme capable of synthesizing DNA in the presence of a DNA template containing suitable primers, and a supply of nucleotides (our A, T, G, and C bases). The DNA polymerase and the nucleotides, directed by the primed DNA template, go to work and after one cycle will have synthesized an extra copy of the double-stranded DNA of our choice, that is we have doubled the amount of our DNA in one step. The process of unwinding the two lots of double strands by heating, then annealing, and copying the four separated strands is repeated, so that after two cycles of PCR, we have four times our original amount of DNA. There is no need to add more primer as it is present in large excess, but what about the DNA polymerase?

We know that proteins in general, and enzymes in particular, are sensitive to heat and often become inactive after such treatment.

The DNA polymerase that was initially used in PCR was isolated from *Thermus aquaticus* (*Taq* DNA polymerase) and it remains active under the heating conditions necessary for PCR. *Thermus aquaticus* is one of the Archaea that live in hot springs, and the biotechnology industry is constantly searching for similarly useful organisms as they sometimes produce such enzymes as the *Taq* DNA polymerase. The inspired idea of Kary Mullis was to recognize that such enzymes would be able to survive the temperatures that are required to separate the DNA double helices formed in one cycle into separate strands, ready for the next DNA polymerase cycle.

The heating/cooling cycles described above for the amplification of DNA have been automated in equipment called thermal cyclers, which are available to molecular biologists. After about 25 thermal cycles, the amplification of our DNA will be about 1 million in practice, usually enough for most purposes, and will have taken just a few hours. Considering the amount of such PCR work now being undertaken, few would deny Kary Mullis his Nobel Prize.

The uses for PCR are several, and the number is increasing steadily. PCR can amplify minute amounts of DNA. By using appropriate primers, small amounts of bacteria and viruses can be detected in tissue samples. This has proven useful in diagnosis. Forensic scientists use PCR routinely, but their collection, handling, and PCR protocols must be extremely robust to avoid contaminating DNA, which can sometimes be a real problem, in the legal sense. Screening for human genetic diseases and amplifying archaeological samples are further useful applications.

The reason for my choice of the PCR to illustrate the application of ancient organisms, via thermal energy cycling, to modern science should by now have become apparent.

GENERAL REFERENCES

Campbell, N.A., Reece, J.B., Urry, L.A., Cain, M.L., Wasserman, S.A., Minorsky, P.V., *et al.* (2008) *Biology*, 8th edn. Pearson Benjamin Cummings, San Francisco.

Garrett, R.H. and Grisham, C.M. (2010) *Biochemistry*, 4th ednBrooks/Cole, Cengage Learning, Boston.

Voet, D., Voet, J.G., and Pratt, C.W. (2002) *Biochemistry*, upgrade edn. Wiley & Sons Inc., New York.