Chapter 2

Introduction: Part II – Genomes, genes and proteins

J. Bryant

School of Biosciences, University of Exeter, Exeter, UK.

Abstract

Changes in DNA provide the variation based on which natural selection works, but increases in the amounts of DNA have also been involved in the evolution of more complex organisms. However, many complex organisms possess much more DNA than is required for their genetic needs; the function of most of the excess DNA is not known. What is known is that in addition to carrying the genetic code, DNA contains within it signals involved in regulating its function. For replication (copying), DNA is organised into units called replicons which contain specific start points (origins). Copying genes into mRNA is regulated by promoters, specific sequences of DNA adjacent to genes. Many genes are active for most of the time but there are also many that, through regulatory mechanisms that operate on their promoters, are responsive to specific cues or signals. In eukaryotic organisms there are further complexities in mRNA synthesis: the initial transcript copied from the gene contains sequences that interrupt the coding sequence. These are known as introns and are removed so that the pieces (exons) of the coding sequence may be joined together. In some genes, there are alternative sites for this splicing process so that one gene may code for more than one protein. Further, there is evidence for movement of introns or exons between genes, generating more variety in evolution. The final product of most genes is a protein; the sequence of amino acids in a protein, coded by the nucleotide sequence in mRNA, determines its shape and hence its function. We cannot yet predict shape from sequence but we are beginning to understand the processes involved in the correct folding of proteins. Finally, amongst the variations in the basic genetic mechanisms, many, perhaps all, organisms can target specific mRNAs for destruction via the synthesis of inhibitory RNA molecules.

1 Introduction

In the previous chapter it was shown that life on earth is centred round a ‘self-replicating’ molecule called DNA which both carries the genetic information from generation to generation and makes

available the information in order to direct, in a biochemical sense, the minute-by-minute activities of each cell. The genetic information is organised as genes, specific linear arrays of deoxyribonucleotides within very long DNA molecules. The sum total of an organism's genetic material, whether we are considering a species or an individual within a species, is called its genome and it is to a discussion of the genome that we now turn.

2 Genome evolution

As was shown in Chapter 1, variations occur in the sequence of individual genes and, further, different combinations of variants exist within a population or species. This genetic variation provides the basis for natural selection and hence evolution. But genomes themselves are also subject to evolution. Consider the figures in Table 1.

Firstly, we note a key division in the classification of living organisms between prokaryotes and eukaryotes. Prokaryotes (bacteria) are simple, usually single-celled organisms in which there are no subcellular organelles. Eukaryotes, which range from single-celled organisms such as yeasts right up to highly complex multicellular organisms, are those organisms in which cells do possess organelles (membrane-bound compartments within cells in which are located the structures and proteins related to particular functions). It is clear that there is some relation between genome size and the complexity of the organism. Prokaryotes have less DNA than unicellular eukaryotes, and amongst the eukaryotes in general, complex organisms have more DNA than simple organisms. This implies that not only can variations be introduced into the DNA sequence during replication

Table 1: Genome sizes (based on compilations in [1–3]).

Species	Taxonomic group	Genome size (base pairs of DNA)
<i>Mycoplasma pneumoniae</i>	Bacteria	1.0×10^5
<i>Escherichia coli</i>	Bacteria	4.2×10^6
<i>Saccharomyces cerevisiae</i> , yeast	Fungi	1.3×10^7
<i>Caenorhabditis elegans</i>	Nematodes (roundworms)	9.7×10^7
<i>Drosophila melanogaster</i> , fruit fly	Insects	1.8×10^8
<i>Fugu rubripes</i> , puffer fish	Fish	3.7×10^8
<i>Xenopus laevis</i> , clawed frog	Amphibia	3.1×10^9
<i>Necturus lewisi</i> , mud-puppy	Amphibia	1.2×10^{11}
<i>Homo sapiens</i> , human	Mammals	3.2×10^9
<i>Arabidopsis thaliana</i> , Thale cress	Plants	1.2×10^8
<i>Vicia sativa</i> , field bean	Plants	3.9×10^9
<i>Vicia faba</i> , broad bean	Plants	2.8×10^{10}
<i>Fritillaria assyriaca</i> , fritillary	Plants	1.2×10^{11}

but also, as mentioned in Chapter 1, on occasions, increases in the amount of DNA may occur. Detailed discussion of this lies outside the scope of this chapter; readers who are interested in the mechanisms that lead to increases in genome size will find a clear account in [1]. All that is needed here is to state that analysis of DNA sequences within and between genomes, combined with extensive observation of DNA replication mechanisms, confirms that genomes can increase in size. With more DNA, there is more 'raw material' for the generation of new genetic variation and hence of new genes.

Intuitively, we might expect this. We could not envisage a vertebrate animal or, for that matter, a plant, functioning with the same number of genes as a bacterium. It seems obvious that the very a complex organism needs more genes than a very simple one. In general, the intuitive expectation is correct but when the detail is examined, the picture is not quite so clear. The first point to make here concerns the actual amounts of DNA. How much DNA does an organism actually need to fulfill its genetic requirements? Certainly not as much as many of the organisms listed in Table 1 actually possess. Indeed, with the exception of the simplest organisms, the DNA amounts are far in excess of the genetic needs, even when allowance has been made for the complexity of the genes themselves and for the DNA sequences involved in regulating genes (such as the promoters mentioned in the previous chapter). This excess of DNA over the likely coding requirements in a genome has been confirmed very clearly for those plants and animals whose genomes have been sequenced. This is known as the C-value paradox (the amount of DNA in the single copy of a genome is the C-value). But it is actually even more paradoxical. Within some groups, the C-value varies little between species, as is seen in birds and mammals. In other groups, e.g. amphibia and plants, there is very extensive variation amongst the species in the group. Why does one plant species need thousands of times more DNA than another? Why does a particular plant or amphibian species need so much more DNA than a mammal? It is a paradox indeed. However, two major points are clear. The first is that, within a group, e.g. flowering plants, many of the larger genomes have arisen by a doubling during evolution of the whole genome. This phenomenon can be induced experimentally in the laboratory, but very occasionally scientists are fortunate enough to come across such an event in nature. This happened, for example, when a genome doubling occurred in a population of an infertile hybrid between two species of *Spartina*, cord-grass (a salt-marsh plant). The genome doubling, which restored fertility, effectively created a new species, *Spartina townsendii*, which has spread extensively round the British coasts from its original location in Southampton Water. Secondly, leaving aside genome doubling, much of the variation in genome size is based not on genes but on sequences that have no direct coding function. Furthermore, many of these non-coding sequences are repeated within genomes; for some sequences this amounts to millions of repeats. So, although the number of potential genes has increased by gene duplication and genetic variation, as more complex organisms have evolved, much of the variation between organisms has been caused by extensive amplification of non-coding sequences. This is beautifully illustrated by comparing two species of bean between which there is a seven-fold difference in DNA amounts (Table 1). These closely related species have very similar numbers of genes; the difference between their genomes is almost entirely based on non-coding repetitive DNA sequences.

The excess of DNA over the apparent coding needs is, in many multicellular organisms, vast, amounting to several orders of magnitude. What is its function? It is not uncommon to hear the excess DNA referred to as 'junk DNA'. However, this is a misleading term. The excess DNA is very unlikely to be functionless 'rubbish'. The synthesis of DNA is, in terms of biochemical resources, a very costly process which makes it improbable that it is undertaken for no reason. The location of a particular type of repeated sequence known as satellite DNA at the centromeres and telomeres suggests a structural function: the telomeres are the ends of chromosomes at which

a special biochemical mechanism exists to prevent chromosomes ‘fraying’ and the centromeres are the structures by which chromosomes are pulled apart during cell division. The involvement of particular DNA sequences in the structural features of chromosomes is another example of the ‘many-sidedness’ of this remarkable molecule. However, for much of the excess DNA in a genome, there is still no clear idea as to its function: some of the mysteries of DNA still remain unsolved.

In organisms possessing the larger genomes, the actual lengths of the DNA molecules are amazing. The DNA of eukaryotic organisms is of course distributed amongst several chromosomes (e.g. 23 pairs of chromosomes in humans), but even so, the individual DNA molecules, one per chromosome, in many multicellular organisms are very long. Indeed, in some plants, some of the DNA molecules are up to a metre long. These are then extraordinary molecules, very long but also very thin: the diameter of a DNA molecule is about 2 nm. A helpful way to consider this is to imagine that the nucleus, the cell compartment or organelle in which the chromosomes are located in eukaryotic organisms, is a ball of 10 cm diameter. On this scale, a DNA molecule of 1 m in length would be magnified to a length of 10 km but would only be 2 mm in diameter.

Even when we shrink the nucleus and the DNA molecules back to their normal sizes, the topological problems imposed by the need to pack very long molecules into the nucleus are huge. This packaging is achieved firstly by coiling the DNA round complexes of proteins called histones. The DNA–histone complex is called chromatin and this is further coiled to form a 30 nm diameter fibre, which, during the process of cell division, becomes very compacted and thus the chromosomes become visible with the light microscope [1]. Those of us who work in this field are amazed that, with this amount of coiling, the enzymes that mediate transcription (see Chapter 1) and replication of DNA can actually locate the correct places along the DNA molecules to start these processes.

3 Organisation of DNA for replication

The basic process of DNA replication, the copying of the genetic information, was described in Chapter 1. However, there is much more to this process than the basic biochemical mechanism. The overall process must be regulated and organised within the context of cell division. The regulatory mechanisms must ensure that DNA is replicated only at the right time within the life of a cell. Description of these mechanisms lies outside the scope of this chapter. Indeed, it would require several chapters to provide a comprehensive description. Here we simply note that regulation involves a network of different sets of mechanisms that operate at different levels, from those that involve direct action on the DNA up to those that link replication with developmental or even environmental cues.

However, one feature of replication that is especially relevant to a consideration of design is the organisation of the DNA molecule for replication. The process of replication does not start just anywhere but at specific sites within the molecule. These sites are known as origins of replication. In the small circular DNA molecules of bacteria, there is one origin of replication and the process proceeds outwards in both directions from the origin (Fig. 1a and b). (Note that this description does not cover the complexity that is inherent in the process that is caused by the two strands of the DNA double helix being in opposite orientations (i.e. are anti-parallel).)

The DNA molecules of bacteria are small and the process of copying an ‘average’ bacterial genome is complete in a matter of minutes. However, the genomes of eukaryotic organisms, and especially those of multicellular eukaryotes, are very different. Firstly, as already noted,

Figure 1: (a) Bacterial DNA molecules are replicated from a single origin of replication. (b) Diagram of the replicative process proceeding from the origin. Reproduced by permission from Bryant, J.A., *Biochemical Mechanisms Involved in Plant Growth Regulation*, eds. C.J. Smith, J. Gallan, D. Chiatante & G. Zocchi, Oxford University Press: Oxford, pp. 139–153, 1994.

the amounts of DNA possessed by eukaryotic organisms are much larger than those in bacteria, sometimes by several orders of magnitude. At the same rate of replication as in bacteria, it would take days rather than minutes to copy these very long molecules if replication started at a single origin of replication.

And it is not only length that poses a problem. As we noted earlier, these enormously long DNA molecules are packaged in an amazing way into the cell nucleus which is only a few micrometres in diameter. The packaging into the individual units of chromatin, the chromosomes, places further constraints on the speed of replication; overall it proceeds, even under optimal conditions, at 100 or more times more slowly than in bacteria. How then do complex multicellular organisms complete DNA replication in hours rather than weeks? The answer is that instead of starting in one place, DNA replication starts in many places along the chromosome (Fig. 2; [4]). This is seen in single-celled eukaryotes such as yeasts, which have a few hundred origins of replication, right through to those species that have very large genomes which possess hundreds of thousands of origins. Thus, as the genome size has increased, the number of replication origins has also increased. Even so, the more complex multicellular organisms take several hours to complete the replication of their DNA.

These replication origins are places along DNA molecules that are recognised by the biochemical ‘machinery’ that replicates DNA. The mechanisms involved in this recognition and in the regulation of the overall process lie outside the scope of this chapter. Nevertheless, to those of us who work on this process, these mechanisms represent beautiful examples of structure–function–control relationships in biological systems. For readers who are interested, some of this is dealt with in [4]. The final point to make here is that flexibility is built into the system. There are times when organisms need to replicate their DNA faster than usual, e.g. in early embryonic growth in many animals. At these times, more replication origins are brought into play such that the amount of DNA replicated from each origin is very much reduced. Thus, not only have eukaryotic organisms acquired, during the course of evolution, multiple start points for DNA replication but they have also developed a built-in flexibility such that start points may be used according to the needs of the organism at any particular stage of its life.

Figure 2: Eukaryotic DNA molecules are replicated from multiple origins of replication, as illustrated here by autoradiographs of replicating pea (*Pisum sativum*) DNA molecules that are incorporating radioactive deoxyribonucleotides at several places along the fibre. The likely position of origins on one of the DNA fibres are indicated by arrows. The scale bar represents 10 μm . Reproduced by permission from Francis, D., Davies, N.D., Bryant, J.A., Hughes, S.G., Sibson, D.R. & Fitchett, P.N., *Experimental Cell Research*, **158**, pp. 500–508, 1985.

4 More on gene structure and function

4.1 Promoters

In Chapter 1 the essential features of gene structure and function were described. The basic features are:

- A promoter ‘upstream’ of the gene, which is essentially the gene’s on–off switch.
- The transcription start and stop sites, which define the stretch of DNA that is copied into messenger RNA (mRNA).
- The coding sequence, which carries the information that instructs the synthesis of a protein during translation.
- At each end of the coding sequence are short stretches of DNA, which although copied in the mRNA molecule are not involved in coding for the protein. These untranslated regions are involved in the mechanics of starting and stopping protein synthesis.

Promoters thus have a key role in regulating the activity of the adjacent gene. How is this done? In general, they do this through the activity of proteins or complexes of proteins called transcription factors. The basic features of the process are the recognition of the promoter sequence by a transcription factor, the binding of the transcription factor to the promoter and making the DNA more accessible to the enzyme (RNA polymerase) that will carry out transcription.

However, a moment's thought will show us that this cannot be the whole story. Even in the simplest bacterium, not all the genes are switched on at any one time. Some genes are obviously needed for the minute-by-minute life of the cell. These are often called 'house-keeping' genes. In genetic terminology, they are constitutively active. However, many genes are switched on or off in response to specific environmental conditions or at particular developmental stages. We can envisage several ways in which this might be done. For example, a negative transcription factor may be bound to a promoter thus keeping a gene switched off except when expression of that gene is required. Alternatively, a positive transcription factor may allow gene expression, either constitutively or under particular circumstances, by binding to the promoter and thus promoting binding and activity of RNA polymerase. Three examples, one from bacteria, one from animals and one from plants, illustrate these principles.

The bacterium *Escherichia coli*, has been for biochemists and geneticists a favourite experimental organism for well over 50 years. In the context of gene expression, it was the organism in which inducible and repressible gene expression was first discovered. That discovery was made by Francois Jacob and Jacques Monod who were later awarded the Nobel prize for their work. The discovery was based on the ability of the bacterial cells to utilise different sugars as a carbon source. In the standard growth medium, the cells are provided with glucose but can actually grow equally well with other sugars. In the classic experiments, cells of *E. coli* were transferred from a growth medium containing glucose to a growth medium containing lactose (the sugar present in milk). Within a few minutes of the transfer, the cells started to make three enzymes that enabled the cells to utilise lactose. The two more important of these are lactose permease (which increases dramatically the rate at which lactose is taken up into the cells) and β -galactosidase (which breaks down the lactose into two smaller sugars, glucose and galactose). The third enzyme, lactose *trans*-acetylase has a minor role in the overall process. What is happening here? First, we note that the genes encoding the three enzymes are activated together. This is an example of *coordinate regulation*. The three genes (named, perhaps confusingly, as the Z, Y and A genes) are actually next to each other in a functional group that Jacob and Monod called an *operon* (this particular operon is known as the *lac* operon). Secondly, the factor that makes this happen is the lactose itself: lactose is therefore an inducer of gene expression. Thirdly, the ability to keep the three genes switched off in the absence of lactose depends on the activity of another gene located next to the *lac* operon. The overall mechanism is illustrated in Fig. 3. The gene whose activity represses the *lac* operon in the absence of lactose produces a repressor protein, i.e. a negative transcription factor that binds to a specific sequence of bases in the promoter. This prevents transcription of the genes. However, when lactose is present, the repressor protein binds to lactose. It has a higher affinity for lactose than for DNA and the binding of lactose changes the shape of the repressor protein so that it cannot in any case bind to DNA. The operon is now available for transcription.

Since this work of Jacob and Monod, further fine-tuning mechanisms operating on this operon have been discovered. However, the key principles remain. Gene expression may be constitutive (as with the gene that encodes the repressor protein) or inducible/repressible (as with the *lac* operon). These principles apply right across the living world, from the simplest to the most complex organism. However, the details vary. Firstly, although coordinate regulation in bacteria is mainly achieved by the organisation of genes as operons under the control of a single promoter, this is not true of eukaryotic organisms, whether they be simple unicellular fungi such as yeast or complex animals or plants. Indeed, genes that are coordinately regulated may be widely spaced from each other, often on different chromosomes. Coordinate regulation is thus achieved because of similar base sequences within their promoters that are recognised by the appropriate positive or negative transcription factors. Secondly, in bacteria many of the inducer substances are small molecules that are either present in the organism's immediate environment, as exemplified by

Figure 3: The induction by lactose of gene expression in the *lac* operon of the bacterium *E. coli*. The three genes of the operon, Z, Y and A, are under the control of a single promoter, P. Adjacent to the operon is a single gene, I, whose promoter allows the gene to be switched on all the time. The protein encoded by the I gene binds to a specific sequence in the *lac* operon promoter and prevents the three genes being copied into mRNA. (Hence, none of the three proteins encoded in the operon are produced.) However, the repressor protein has a higher affinity for lactose than for the DNA sequence in the promoter and thus when lactose is present, the repressor protein leaves the *lac* operon promoter and binds instead to lactose. The *lac* operon genes are now available for mRNA synthesis, leading to the production of the proteins that metabolise lactose. Note that the diagram is not to scale (otherwise the lactose molecule would be too small to be visible).

lactose, or generated during cellular metabolism (in which case they may function to induce or repress the expression of enzymes involved in particular metabolic interactions or pathways). In eukaryotes, gene induction or repression by metabolites certainly occurs and indeed is common in unicellular eukaryotes such as yeast. However, in more complex eukaryotes, induction or repression of gene expression is very much more a key part of development than it is in unicellular organisms, which means that developmental cues are very important in gene regulation. In addition, there are also many examples of responses at the gene level to environmental signals, some of which may interact with developmental cues. These principles of developmental and environmental regulation are illustrated by two well-known examples, namely the regulation of gene activity by steroid hormones in mammals and by light in plants.

Different steroid hormones regulate a range of developmental, growth and maintenance processes in mammals. Their action on gene expression is in general to act as an inducer, but they do not interact directly with the relevant genes. When a steroid hormone such as testosterone enters a cell that is competent to respond, it encounters a steroid-binding protein or hormone receptor. The receptor proteins with their bound hormone molecules then form pairs – dimers; the dimerised

....**GGA/TACANNNTGTTCT**.... →
CCT/ATGANNNACAAGA....

Figure 4: The androgen (male hormone) response element, the sequence in the promoters of androgen-responsive genes to which the hormone receptor protein plus hormone binds. The gene is 'downstream' from this sequence, as shown by the arrow.

receptors travel to the cell nucleus and bind to the gene promoters that contain a specific sequence called the androgen response element (Fig. 4). The genes associated with these promoters are thus coordinately induced even though they are spread throughout the genome.

In plants, light is a major regulator of developmental processes, as well as being the energy source for the process of photosynthesis. The effects of light on developmental processes are mediated via a number of light receptor molecules and involve changes in gene expression. The most abundant of these light receptors are the phytochromes which exist in two states. In one of the states, P_r , it is receptive to light in the red region of the spectrum with a peak wavelength at 660 nm. The light energy causes a change in the shape of the non-protein part of the molecule (see Fig. 5a) and in most of the processes involving the phytochrome system it is this form, known as P_{fr} , that is active. The active phytochrome is then able to trigger a range of molecular, cellular or developmental responses, according to the specific situation in which it has been activated. It may also be inactivated by one of three routes (Fig. 5b). Firstly, under the influence of far-red light (peak active wavelength 730 nm), P_{fr} is converted back to its inactive form, P_r . Secondly, this conversion also takes place slowly in the dark. Thirdly, P_{fr} can be slowly broken down, probably by a specific protein-hydrolysing enzyme (protease).

Thus, cells are equipped with sensing systems, many of which are involved in the induction or repression of gene activity. This relies on specific transcription factors and on specific sequence elements within promoters and leads to the controlled expression of genes in response to internal or external conditions or at particular developmental stages. Of necessity, this brief description cannot convey the amazing complexity of gene regulatory mechanisms that exist across the living world. A full description of what we know about these (and there is much that we still do not know) would take a large volume of its own. However, in order to convey some of the wonder of this, let the reader imagine the gene control mechanisms involved in the development of an adult human, consisting of around 70 million million cells (and over 200 different types of cells), from a single cell, the fertilised egg.

4.2 mRNA synthesis in eukaryotic cells

In the previous chapter the basic mechanisms involved in bacterial gene expression were presented. However, in eukaryotes, there are a number of further aspects that we now need to consider. The first of these is the addition of extra residues at each end of the newly synthesised mRNA molecule. At the 5' end, the end that engages first with the ribosome (Fig. 1 in the previous chapter), a single modified GTP is added in the reverse orientation to normal, making a 5' to 5' triphosphate bridge. This structure is known as the mRNA cap and facilitates the binding of the mRNA to the ribosome. At the 3' end, a series of adenosine monophosphate residues is added. This poly(A) tail is a stability regulator. For example, the mRNA that encodes globin (the protein part of haemoglobin) has a poly(A) tail of about 200 residues and may remain functional for up to 2 weeks in human red blood cells. At the other extreme, the mRNAs encoding the various histones, the major proteins

Figure 5: (a) The non-protein part of the plant light receptor phytochrome changes its configuration according to the wavelength of the light it perceives. The diagram shows the ‘flip’ from one configuration to the other and also indicates the attachment to the protein part of the molecule. (b) Interconversions between the different forms of phytochrome. Reproduced by permission from Raven, P.H., Evert, R.F. & Eichhorn, S.E., *Biology of Plants*, 7th edn, Freeman: New York, 2005.

Figure 6: Most eukaryotic genes contain introns, non-coding DNA sequences that interrupt the coding regions. These are copied into pre-mRNA which is immediately modified by the addition of the cap (a modified base) and the poly(A) tail. The introns are then removed by a cut-and-rejoin mechanism in order to obtain finally the uninterrupted mRNA sequence.

associated with DNA in the chromosomes, are only needed for a few hours during the process of DNA replication. These mRNAs have no poly(A) tails.

In a sense, these additions are not remarkable. They are simply reminders that despite the universal genetic code and the general similarities of gene function across the living world, there are variants in the details of the mechanisms themselves and their regulation in different types of organism. However, another of the differences between eukaryotes and prokaryotes, when initially discovered in the 1970s, met with a degree of surprise that bordered on disbelief. The discovery was that many eukaryotic genes (in complex multicellular eukaryotes, most of their genes) contain sequences of DNA which interrupt the coding sequence (Fig. 6) and that these sequences are copied during mRNA synthesis. These interrupting sequences are removed by an editing process after the addition of the cap and the poly(A) tail. Thus, in eukaryotes, we refer to the unedited version as pre-mRNA; the interrupting sequences in the genes are known as *introns* and the coding sections are called *exons*.

4.3 Splicing and shuffling

The removal of the introns from the pre-mRNA must of course be a highly regulated process in order that the correct coding sequence is retained in the mRNA. It will not surprise readers to know that once again, recognition of specific base sequences is involved. The splicing out of the introns is carried out by complexes of proteins and small nuclear RNAs (snRNAs). The snRNAs bind to the splice sites by specific base pairing, enabling the splicing enzyme to cut the RNA at that site. In addition, a looping out of the intron brings the two cut sites at either end of the intron into close proximity so that the cut ends may be joined. The mRNA is thus 'edited'.

In encountering such mechanisms a biologist will always ask why it has evolved. As with the large excess of DNA in many eukaryotic organisms, it seems so wasteful. However, within the mechanism just described there are possible indicators of evolutionary advantage. The first of these is that in many genes, alternative splicing sites exist. That is, whereas particular sites may be preferred, others may, under specific conditions, be used, thus generating alternative mRNAs

from an individual gene. Whilst this may seem strange, set against the large excess of DNA, it is not so strange if we consider that it may be advantageous to have two proteins whose synthesis is under the control of a single promoter region.

Secondly, by examining the sequences of genes encoding proteins that have some properties in common, it appears that the exon-coding regions may recombine between genes, thus creating new genes encoding new proteins. Readers wishing to know more about this topic will find a clear and concise account in [1].

5 Gene sequence and cellular function

The essential working molecules of cells are the proteins. As described in Chapter 1, proteins are made of building blocks called amino acids, of which there are 20. It is the order of amino acids that is encoded in the DNA and the order of amino acids that governs the overall shape of a protein. This is because the different amino acids, in addition to being joined end-to-end to form the ‘backbone’ of the protein, can also interact with each other, forming linkages of various types between different parts of that backbone. The protein thus folds up into a specific 3D shape. That specific shape may appear to the non-expert as a ball of tangled string, but it is actually vital for the protein’s function. Thus, the ability to mediate specific chemical reactions (i.e. to act as an enzyme) or to perform a structural function or to form a channel for taking metal ions into the cell, or indeed any other function is dependent on the 3D shape of the protein.

As already noted, this 3D shape depends on the order of particular amino acids within the protein chain and their ability to interact with amino acids elsewhere in the chain. In a very real sense then the shape of the protein is built into the genetic code. It is almost the ‘holy grail’ of protein biochemistry to be able to predict from the DNA coding sequence (from which we know the order of amino acids) what the 3D shape of a protein will be. However, this has eluded all the best efforts of biochemists and bioinformaticists. Nevertheless, we are beginning to understand, thanks to the elegant work of, amongst others, Professor Christopher Dobson, the mechanisms involved in a protein ‘getting into shape’ [5]. The final structure of the protein is in general the most thermodynamically stable, the configuration of least energy. However, the number of possible interactions between amino acids is very large, especially in large proteins. It is apparent that during the folding process some interactions form the core elements in the final shape and that around these the protein ‘tries’ different configurations until the final shape is achieved. We should note that the ability to do this involves the reversibility of amino acid interactions and that the less stable interactions are those that are more readily reversed. The final product of this process is a protein that is exquisitely fit for its function. Indeed, the distinguished American biochemist, Dan Koshland, has referred to enzymes as beautiful, finely tuned miniature machines. One amazing example of this is the enzyme ATP synthase which carries out the final step in trapping the energy from respiratory oxidation. This complex enzyme consists of several individual protein molecules, some of which rotate like a tiny turbine, operating at a scale of nanometres. One can only react with wonder, even awe, at such miniature elegance!

6 How many genes are needed?

We have already noted that the amounts of DNA in many, perhaps most, multicellular organisms appear to exceed, often by several orders of magnitude, the amount needed to satisfy the genetic

requirements of an organism. But how many genes does an organism actually need? It is actually quite difficult to obtain an accurate answer to this question but 'ballpark' figures, based on genetic analysis and on biochemical data, are of the order of 25–30,000 for plants and perhaps 80–100,000 for mammals. However, the advent of genome sequencing provides a much more direct approach to the problem: find all the genes and count them. When this is done, most of the data confirm our view that increased biological complexity requires greater numbers of genes. However, the data for mammals are totally surprising. Instead of 80–100,000 expected, humans and other mammals possess only 25–26,000, around the same as a flowering plant. Birds, also complex warm-blooded animals, have 20–23,000 genes. Mammals and birds are clearly much more complex than flowering plants and yet manage with about the same number of genes. However we view this, it is clear from comparison of gene numbers with genetic function that birds and mammals do not seem to have nearly enough genes.

At least a partial answer to this conundrum is that some genes, perhaps a significant proportion of genes, are multifunctional. We have already noted that by splicing a pre-mRNA in different ways it is possible to generate more than one type of mRNA. It is also possible to envisage alternative start sites within genes, and we certainly know of ways in which particular proteins may be processed or modified in different ways after they have been formed. All these mechanisms will increase the number of gene products that are encoded within a set of genes. This provides an efficiency of use of DNA sequences that had previously been encountered only as a few rare examples.

However, these solutions to the gene number paradox in mammals again raise the question of the excess DNA. If mammals can exist with only around 25–35% of the expected gene numbers, then the amount of DNA that is in excess of coding needs is even greater. It is easy to understand the selective pressure to maximise the coding usage of DNA in very small genomes such as those of viruses. It is much harder to understand what the selective pressure has been to maximise the use of individual genes in organisms that appear to have no need to do this.

7 Variations on a theme

In this and the previous chapter I have set out the main ways in which genomes are replicated and genes are regulated. Most genomes consist of DNA and at the heart of the systems that replicate them are the enzymes that copy the template DNA to make the daughter molecules. These are the *DNA-dependent DNA polymerases*. However, in many viruses, the genome is made not of DNA but of RNA. There are two main ways in which these genomes are replicated. First, in the retroviruses, the RNA genome is copied into a DNA version by an *RNA-dependent DNA polymerase*, often referred to colloquially as reverse transcriptase. This DNA version of the genome serves as a template for making new RNA copies using *DNA-dependent RNA polymerases* (similar to the enzymes that make the 'normal' types of RNA in a cell). Some of these retroviruses insert the DNA copy of their genome into the host's genome thus leaving a permanent copy of the virus's genes. An example of this is the human immunodeficiency virus (HIV), the causative agent of acquired immunodeficiency syndrome (AIDS). Secondly, in many RNA viruses, and particularly in those that infect plants, the viral RNA genome is copied by an *RNA-dependent RNA polymerase*. A summary of these various polymerases is shown in Table 2.

Mention of this last group of enzymes, the RNA-dependent RNA polymerases, brings us to consider a very exciting recently discovered group of mechanisms for regulating RNA populations within cells. In the 1970s and early 1980s, investigators studying the infection of plants by RNA viruses such as tobacco mosaic virus were not surprised to find that infected plants contained

Table 2: Enzymes that synthesise nucleic acids in a template-dependent manner.

Enzyme	Template	Product	Notes
DNA-dependent DNA polymerase	DNA	DNA	Used in replicating and repairing DNA
DNA-dependent RNA polymerase	DNA	RNA	Makes RNA copies of genes and other coding sequences
RNA-dependent DNA polymerase	RNA	DNA	Involved in replication of some types of virus
RNA-dependent RNA polymerase	RNA	RNA	Involved in replication of some types of virus; also involved RNA-dependent gene silencing

RNA-dependent RNA polymerases that were involved in replicating the viral genome. However, what was surprising was that uninfected plants were often shown to possess RNA-dependent RNA polymerases of their own, the function of which was completely unknown [6, 7]. We now know, firstly, that double-stranded RNA molecules are targets in plants and animals for a specific degradation pathway. In plants this provides one of the defences against RNA viruses that must go through a double-stranded RNA phase, at least transiently, every time a virus RNA genome is copied. However, more intriguingly, plant and animal cells can target specific mRNAs for this destruction pathway by copying them with RNA-dependent RNA polymerases, thus making them double-stranded. (Thus we at last know why uninfected cells have this type of polymerase.) Further, the ability to target specific messages means that the activity of a particular gene may be quickly silenced, giving rise to the term ‘RNA-based gene silencing’. The details of the mechanism for destruction of the double-stranded RNAs lie outside the scope of this chapter. Here we just need to note that the pathway involves the generation of small RNA molecules which are themselves very inhibitory and which, by careful design of the RNA sequence, can be used by scientists to silence individual genes. These small RNA molecules are known as small inhibitory RNAs and the general mechanism as RNA interference (RNAi). Any reader who is interested in finding out more will find clear accounts of this exciting topic in [8] and [9].

8 Concluding remarks

These first two chapters have outlined the key features of information maintenance, transmission, retrieval and use that lie at the heart of life on earth. They have also outlined some of the ways in which these processes are regulated, providing food for thought for anyone who turns to nature for inspiration about regulation of human industry and commerce. However, it is apparent, even from this overview, that there is much that we do not yet know (and some things that we probably never will know). Nevertheless, we are beginning to understand some of mechanisms involved in the way in which these core processes are used in the life of the whole organism. Much of the rest of this book is concerned with this topic, as we examine in more detail, in specific situations, the functioning of genes and proteins before finally moving on to considering some of the higher order functions of particular living organisms.

References

- [1] Brown, T.A., *Genomes*, 2nd edn, Bios: Oxford, 2002.
- [2] Bryant, J.A. (ed.), *Molecular Aspects of Gene Expression in Plants*, Academic Press: London, 1976.
- [3] Gregory, T.R., <http://www.genomesize.com/>, last consulted on 12/12/2005.
- [4] Bryant, J.A., Moore, K. & Aves, S.J., Origins and complexes: the initiation of DNA replication. *Journal of Experimental Botany*, **52**, pp. 193–202, 2001.
- [5] Dobson, C.M., Protein folding and misfolding. *Nature*, **426**, pp. 884–890, 2003.
- [6] Evans, D.M.A., Bryant, J.A. & Fraser, R.S.S., Characterization of RNA-dependent RNA polymerase activities in healthy and TMV-infected tomato plants. *Annals of Botany*, **54**, pp. 271–281, 1984.
- [7] Evans, D.M., Fraser, R.S.S. & Bryant, J.A., RNA-dependent RNA polymerase activities in tomato plants susceptible or resistant to tobacco mosaic virus. *Annals of Botany*, **55**, pp. 587–591, 1985.
- [8] Black, D. & Newbury, S., RNA interference – what it is and what it does. *The Biochemist*, **26**, pp. 7–10, 2004.
- [9] Zhang, J., Special delivery – small RNAs silencing gene expression. *The Biochemist*, **26**, pp. 20–23, 2004.