1

Gene organisation and control

The purpose of this introductory chapter is to present a brief overview of the organisation, packaging, transcription and regulation of genes in animal systems. In so doing, neither detailed explanations nor supporting evidence will be offered, although a few key references will be given where appropriate. It is assumed that much of the material summarised below will be familiar to readers from courses in molecular biology, and that further detail would be largely superfluous. Further explanation of particular points may be sought from any of several excellent texts covering the cell and molecular biology of both pro- and eucaryotes, including Lewin (1987), Watson (1987), Alberts *et al.* (1983) and Darnell *et al.* (1986).

Some knowledge of the following molecular techniques will be assumed (without further explanation) in the rest of the book: (i) DNA reannealing and RNA/DNA hybridisation for studying populations of nucleic acids; (ii) basic cloning technology for the isolation and bulk preparation of particular DNA sequences; (iii) the construction of chimaeric genes in which regulatory elements from one gene are fused to a foreign coding sequence (often a procaryotic reporter gene whose product can easily be detected); (iv) Southern blotting and restriction mapping with a cloned probe in order to study the genomic organisation of a gene; (v) basic DNA sequencing (Gilbert/Maxam and Sanger) techniques; (vi) Northern blotting to assess the size, tissue distribution and quantitative expression of transcripts from any gene for which a cloned probe is available; (vii) *in situ* hybridisation with a cloned probe to localise specific RNA transcripts; (viii) immunodetection with poly- or monoclonal antibodies to localise particular proteins; (ix) Western blotting to identify and quantify specific proteins (after electrophoretic separation) using appropriate antibodies.

1.1 DNA organisation

All eucaroytes contain vastly more DNA per haploid genome than can be accounted for by the coding genes, of which there are between 4000 and about 100 000 different types in various animals. A variable proportion of the DNA is repetitive, consisting of multiple copies of the same or very similar sequences. Tandem arrays of short simple sequences repeated (inexactly) millions of times per haploid genome constitute the 'satellite' or simple-sequence DNA fraction, which reforms duplex structures rapidly during the reannealing of denatured DNA. Simple sequence DNA serves no coding function, and is transcribed only during oogenesis (see § § 2.3 and 2.4.2). Although such DNA may play some role in the structural organisation of chromosomes, much of it may be non-functional – sometimes described as 'junk' DNA or evolutionary debris. Other pointers towards this view of much of the eucaryotic genome include the frequent occurrence of transposable elements (see below) and of 'pseudogenes'; these latter are variant versions of *bona fide* genes which have become mutated in various ways so as to preclude their expression (e.g. non-functional promoters and/or premature stop codons: see § § 3.2.4).

By contrast, some coding genes are included in the 'middle-repetitive' DNA fraction present usually in a few hundred to a few thousand copies per haploid genome. These include the major ribosomal genes (encoding the 18S + 5.8S + 28S rRNA species), the 5S rRNA genes and the histone genes in, for example, sea urchins (α-type; see § 2.4.2) or *Drosophila*. These genes are arranged as tandem (head-to-tail) clusters each containing many copies of the same repeat unit, which normally includes the coding gene(s) plus 'spacer' sequences. Other repetitive genes show less regular organisation, with copies either loosely clustered (e.g. the tRNA genes, or histone genes in mammals) or dispersed to many different sites in the genome. Frequently, the different copies are variations rather than exact repeats, intergrading into families of related genes which have apparently diverged from a common ancestor. Examples of such genes include the actin genes (§ 3.1), the vertebrate globin genes (§ 3.2.4) and the vitellogenin genes (§ § 3.3.3 and 4.2). Some dispersed repetitive sequences are mobile and can transpose from one chromosomal site to another; there is evidence that some of these elements may be related to retroviruses. Mobile genetic elements often cause mutations at their sites of insertion, and this can be exploited in various ways by the experimenter (e.g. the Tc1 transposon in *Caenorhabditis elegans*; see § 4.2). Finally, transposable

elements can be used as vectors for introducing novel genetic constructs into every cell of an embryo, as in the case of *Drosophila* P elements which only become transposed in the germ line (see § 5.1.3).

Sequences present in one or a few copies per haploid genome are described as unique or single copy. The vast bulk of protein-coding genes fall into this category. Because single-copy sequences reanneal slowly during DNA renaturation, it is possible to isolate them as that fraction which remains single-stranded after all repetitive sequences have reformed duplex structures; these latter bind to hydroxyapatite, whereas single-stranded DNA does not. Preparations of single-copy DNA (as a labelled tracer) are especially useful when comparing different populations of RNA by means of RNA-excess hybridisation (see § 2.3). If total DNA is used, hybrid formation between repetitive genes and their transcripts tends to obscure the much slower reaction between single-copy genes and their corresponding RNAs. In passing, we may note that most of the genes with which we shall be concerned in this book (i.e. protein-coding genes with important functions during animal development) fall into the single-copy category. However, this does not mean that repetitive DNA is unimportant; multicopy sequence elements located within coding or regulatory regions are often shared by functionally related groups of genes, and these have provided important clues about their modes of action and/or control. Two recurrent examples of this are the 'zinc-finger' and 'homeobox' elements which encode DNA-binding protein domains (see § 5.6).

Much of the genome in most eucaryotes is organised in a short-period interspersion pattern typified by the *Xenopus* genome, where short repetitive sequences (averaging 300 bp in length) alternate with longer single-copy sequences (averaging 750 bp). However, in *Drosophila* and certain other insects, the genome is organised in a long-period interspersion pattern, with extended blocks of repetitive DNA (averaging 5600 bp) separated by vast stretches of unique sequences (at least 13 000 bp in length). The significance of this contrast remains unexplained, since other insects including some dipterans (e.g. *Musca*) show *Xenopus*-type sequence interspersion, as do other organisms with very small genomes such as *Caenorhabditis elegans* (see § 4.2).

1.2 Chromatin and DNA methylation

In eucaryotic nuclei, the DNA is associated with a variety of nuclear histone and non-histone proteins in the form of **chromatin**. The fun-

damental repeating structure of chromatin is the nucleosome, which recurs at approximately 200 bp intervals along the length of the chromosomal DNA. Within each nucleosome, 145 bp of core DNA is wound around a histone octamer comprising two molecules each of the H2A, H2B, H3 and H4 core histones (this DNA–histone complex is termed the core particle). A single molecule of the fifth major histone, H1, is bound at the edge of the nucleosome to the 60 bp of linker DNA between adjacent core particles. A variety of nuclease enzymes cleave chromatin at approximately 200 bp intervals by making double-stranded cuts within the linker DNA. Incomplete digestion of chromatin with such enzymes (e.g. micrococcal nuclease) produces a 'nucleosome ladder' of DNA fragments with lengths of about 200 bp, 400 bp, 600 bp, etc., representing nucleosome monomers, dimers, trimers, etc. Under certain conditions, the ends of linker DNA flanking each cut site may become digested away, reducing the monomer fragments to 145 bp of core DNA protected by association with the histone octamer.

Nucleosomal chromatin is relatively resistant to digestion by the enzyme DNaseI; however, in the neighbourhood of active or potentially active genes, the chromatin is approximately 25-fold more sensitive to DNaseI attack. Thus the chick ovalbumin gene is DNaseI-resistant in brain chromatin where it is inactive, but much more sensitive in laying hen oviduct where it is highly active (see Garel & Axel, 1976 and § 3.3.2). EM spreading shows bulk chromatin as a characteristic 'string of beads' (nucleosomes evenly spaced along the DNA axis), but as a smooth fibre in highly active gene regions. These latter are also distinguished by the presence of side branches representing nascent RNA transcripts. However, nuclease digestion experiments indicate that a 200 bp periodicity is retained within DNaseI-sensitive smooth-fibre chromatin; moreover, all four core histones are present, as shown by immunological studies. These features suggest that DNaseI-sensitive 'active' chromatin represents a more extended conformation of the basic nucleosome structure. Whereas the extended DNA length is compacted by some 6–7 fold in nucleosomal chromatin (through being wound twice around each nucleosome), the degree of compaction is only about 2-fold in smooth-fibre chromatin. Another distinction between active and nucleosomal chromatin is the presence of two specific non-histone proteins (HMGs 14 and 17) bound to the linker DNA, possibly replacing histone H1 (see Weisbrod, 1982). Although most actively transcribed genes are in the smooth-fibre DNaseI-sensitive conformation, the converse is not necessarily the case. Thus some

inactive genes are found to be DNaseI-sensitive, usually in tissues where they either have been or will become active (see § § 3.2.2 and 3.3.2). H1 may act as a general transcriptional repressor in nucleosomal chromatin, as well as being involved in the higher-order packing of chromatin into supercoils, etc.

Within broad regions of DNaseI-sensitive chromatin, certain limited sites are found to be especially susceptible to DNaseI attack. These DNaseI-hypersensitive (DHS) sites arise where particular protein factors bind to target sequences in the DNA (see e.g. Elgin, 1984). In many cases, DHS sites are found close to the 5′ ends of actively transcribed genes, but other DHS sites may occur well upstream from, within or downstream from such genes. The presence of DHS sites in a given gene region can be demonstrated by treating chromatin with extremely low concentrations of DNaseI and then digesting the DNA completely with an appropriate restriction enzyme. Among the resultant DNA fragments which hybridise with the corresponding gene probe will be a set of DNaseI-generated sub-bands as well as the standard restriction bands for that enzyme. The precise locations of DHS sites can then be mapped using indirect end-labelling methods. In cases where the same gene is active in different tissues and/or at different levels, alternative sets of DHS sites may characterise that gene region in each of its activity states (see Fritton *et al*., 1984; § 3.3.2). Finally, there is evidence that certain DHS sites may 'mark out' a gene for future activation, and that such patterns of DHS sites may be inherited over many cell generations without actual expression of the gene (Groudine & Weintraub, 1982). Such a mechanism might perhaps underlie the phenomenon of determination, whereby sets of tissue-specific genes are selected for activation well before their products become expressed in detectable amounts.

Yet another possible explanation for determination and for the stability of the differentiated state (both heritable characteristics) involves the pattern of DNA methylation. Vertebrate DNA is typically methylated at a high proportion of C residues occurring in CG dinucleotides. Note, however, that there is no DNA methylation in many invertebrates, including *Drosophila* (Urieli-Shoval *et al*., 1982). Methylated C residues are not incorporated directly into newly synthesised DNA strands. Rather, the methylation of C residues in vertebrate DNA is accomplished after replication by a methylase enzyme which is specific for (i) Cs within CG dinucleotides and (ii) hemimethylated DNA (Gruenbaum *et al*., 1982). Because of semiconservative DNA replication, each daughter duplex will be hemimethylated, i.e. will contain one

methylated parental strand and one newly synthesised non-methylated strand. The methylase will thus add methyl groups only to those Cs occurring in CG dinucleotides on the new DNA strand at positions where neighbouring Cs on the opposite parental strand are already methylated. In this way, the parental pattern of methylated and unmethylated sites will be passed on exactly from parent to both daughter cells, and so on. Methylation patterns in a particular gene region can be analysed by using two isoschizomer restriction enzymes which recognise the same target sequence containing a CG dinucleotide (e.g. C*CG*G, but which are respectively sensitive (*Hpa*II) and insensitive (*Msp*I) to methylation of the central Cs. Sites cut by *Msp*I but not by *Hpa*II identify some of the methylated Cs (only those contained in C*CG*G target sequences) within the DNA region mapped, whereas unmethylated target sequences will be cut by both enzymes.

There is considerable evidence that many active genes, and especially their 5′-flanking regions, are relatively undermethylated compared with bulk DNA, suggesting that demethylation might play some role in gene activation in vertebrates. However, this correlation may be coincidental, or else demethylation may be a consequence rather than a cause of transcription. The evidence on this point remains equivocal (Bird, 1984, 1986; Cedar, 1988). Demethylated CG-rich sequences are found near many housekeeping genes that are active at low levels in all cell types, but they are often absent from the neighbourhood of tissue-specific luxury genes (Bird, 1986). Thus demethylation is unlikely to play a general role in activating these latter, although it may mark out the former for constitutive activity. Methylation of DNA target sequences can prevent them from being bound by transcription factors (Iguchi-Ariga *et al*., 1989), and there is also a mammalian protein which binds specifically to DNA containing methylated CGs (Meehan *et al*., 1989).

Recently, differential methylation of maternal (low) and paternal (high) genomes during vertebrate gametogenesis has been implicated in genomic imprinting, whereby the two parental alleles of a gene may show differential activity during development of the zygote. However, the overall picture is complicated both by a loss of methylated sites during early development and later by *de novo* methylation of many new sites (except in the germ line, where previous imprinting may be erased during meiosis). Differential methylation may not be the sole mechanism involved in imprinting, and may only maintain the inactive state of non-expressed DNA (Monk, 1988). Heritable chromatin structures (e.g. patterns of DHS sites) may also be involved in the

imprinting phenomenon, just as both may help to propagate active versus inactive chromatin states for particular genes through many cell generations.

Before leaving these general features of active chromatin, it is worth noting that active genes are found associated with the nucleoskeleton or nuclear cage (see Jackson *et al.*, 1981; Ciejek *et al.*, 1983; Jackson & Cook, 1985) at or near the periphery of the nucleus (Hutchison & Weintraub, 1985). Bulk nucleosomal chromatin is transcriptionally inactive (repressed) through association with histone H1. By contrast, the DNA in active gene regions appears to be under torsional strain (see review by Weintraub, 1985), and is associated with topoisomerase I (see e.g. Gilmour *et al.*, 1986) which may function in DNA supercoiling.

1.3 Transcriptional control

1.3.1 RNA polymerases and their products

There are three nuclear RNA polymerase enzymes in eucaryotes; these differ in their (complex) subunit compositions, their sensitivities to the fungal toxin α-amanitin, and in the sets of genes which they transcribe. Polymerase I (pol I) is resistant to α-amanitin and localised in the nucleolus. Its sole function is to transcribe the rDNA repeats into long (7–12 kb) rRNA precursors, each including one copy of the 18S, 5.8S and 28S units. These structural rRNA sequences become methylated and are cleaved out post-transcriptionally from the long precursor; the remaining (non-methylated) parts of this precursor are discarded during processing. All of this occurs within the nucleolus, where mature rRNAs are then assembled into ribosomes.

Pol III is inhibited by high but not low concentrations of α-amanitin; it is found throughout the nucleoplasm, and transcribes a limited set of small stable RNA species, including the 5S rRNAs, all tRNAs, and the U6 small nuclear RNA (snRNA; see below). Whereas 5S RNA requires only limited end-trimming prior to export, tRNAs are subjected to extensive post-transcriptional processing – including many unusual base modifications and often the removal of a small intron sequence adjacent to the anticodon loop. The process of intron removal from tRNA precursors follows a pathway distinct from that involved in pre-mRNA splicing (to be discussed in § 1.4 below). Some protozoans have intron-containing rRNA genes, and in such cases the intron sequences are removed from rRNA precursors by yet a third

distinct mechanism (self-splicing, i.e. not requiring protein or other RNA cofactors; see Cech, 1983).

The remaining RNA polymerase, pol II, is also nucleoplasmic in location, but is sensitive to very low concentrations of α-amanitin. It transcribes all protein-coding mRNA precursors as well as most of the U-series snRNAs and a variety of transcripts with undefined functions. These last often include interspersed repetitive sequences, are apparently non-translatable, and are nucleus-confined in most tissues apart from oocytes (see also § § 2.3 and 2.4.2). Because the majority of protein-coding genes contain introns (with some exceptions, e.g. most histone genes), the removal of these sequences by the splicing of primary transcripts is an essential step in generating functional mRNAs for export to the cytoplasm. Nascent pol II transcripts become modified at their 5′ ends by capping, whereby a G residue (from GTP) is linked in reverse orientation through a triphosphate bridge onto the true 5′ terminus of the RNA (termed the cap site). This reversed G subsequently becomes methylated, and the resultant cap structure apparently protects the RNA from degradation. After the completion of transcription, a series of 50–200 successive adenosine residues [poly(A)] is added onto the 3′ terminus of most pre-mRNAs (again, the major histone mRNAs are exceptional on this count). Both splicing and polyadenylation will be covered in more detail in § 1.4 below. However, we may note in passing that the 3′ poly(A) tag provides a convenient means for purifying polyadenylated mRNAs and their precursors from the bulk of non-polyadenylated r- and tRNAs, since the former but not the latter will bind to an oligo(dT)-cellulose column.

1.3.2 Promoter sites and other transcriptional control elements

A number of techniques are available for identifying essential promoter elements and other sequences needed for optimal transcription, as well as for locating the sites where proteins bind to such DNA sequences. The latter are most often studied by variations on DNA footprinting or nuclease protection assays, whereby proteins bound tightly to sites within a DNA fragment will protect those sites from nuclease degradation; the sequences of both protected and non-protected regions are then determined (separately) to establish the nature and location(s) of the binding site(s). In the former category, DNA elements important for transcription can be identified by comparing the levels of gene expression obtainable from different lengths of regulatory sequence; usually this means the 5′-flanking regions of the gene,

but sometimes includes intragenic and/or 3′-flanking sequences. Cloned genes plus their flanking regions may be transcribed *in vitro*, injected into *Xenopus* oocyte nuclei, or introduced by various means (microinjection, transfection or cotransformation) into heterologous cell types. Some of these approaches can be criticised on the grounds that tissue-specific transcription factors may be lacking in the recipient cell type. However, these should be present in nuclear extracts derived from the appropriate tissue, as used for *in vitro* transcription assays. If, on the other hand, cloned genes are introduced into homologous cells, background expression of the endogenous gene copies may pose problems. These are most simply circumvented by fusing regulatory regions from the gene of interest onto the coding region of a foreign 'reporter' gene. This last can encode any product which is never expressed by the host cells and for which there is a convenient assay. Typical choices for animal systems include the procaryotic genes encoding neomycin resistance (a selectable marker), chloramphenicol acetyltransferase (whose activity is easily quantitated), or β-galactosidase (whose protein product can be detected *in situ* by histochemical staining).

Various methods have been devised for introducing normal or fusion genes into every cell of a developing organism, including the germ line. Only thus is it possible to determine whether the flanking sequences used (from gene X) are sufficient to endow the introduced gene with the same pattern of temporal, spatial and quantitative regulation that characterises the expression of X during normal development. Techniques for achieving this in *C. elegans* are mentioned in § 4.2, and the pioneering approach using transposable P elements in *Drosophila* is discussed in § 5.1.3. Transgenic mice can be created by fertilising eggs with sperm *in vitro*, microinjecting gene constructs into the male pronuclei, and then transferring the early embryos into the wombs of pseudopregnant mothers; some of these embryos develop to term and express the introduced gene with varying degrees of specificity (see review by Palmiter & Brinster, 1985). At least in some cases, the introduced genes are expressed only in appropriate tissues in transgenic mice, for instance the rat elastase I gene in the exocrine acinar cells of the pancreas (Swift *et al.*, 1984), or the rat myosin light-chain 2 gene in skeletal muscle cells (Shani, 1985). Variations on this method can be used, for example, to target the expression of reporter genes by linking them to tissue-specific regulatory sequences; thus a rat elastase I promoter fused to a human growth hormone gene is specifically expressed in pancreatic acinar cells (Ornitz *et al.*, 1985). Similar constructs involving a diphtheria-toxin reporter gene destroy the

developing pancreatic acinar cells, because only these cells express the toxin (Palmiter *et al*., 1987). On occasion, a host gene may become disrupted fortuitously through the integration of introduced sequences, causing a novel mutant phenotype in that transgenic mouse; by cloning the DNA sequences flanking such inserts one can identify the wild-type gene (see Woychik *et al*., 1985). Finally, it is possible to cure specific genetic defects by introducing wild-type copies of the appropriate gene into mouse zygotes of the mutant strain; thus complete myelin basic protein (MBP) genes can rescue the homozygous *shiverer* defect (involving partial deletions of both endogenous MBP genes: Readhead *et al*., 1987). Such a cure extends even to the germ line (since *all* cells carry wild-type MBP genes) and is thus inherited by offspring of such transgenic mice.

The three eucaryotic RNA polymerases recognise different sequence elements in the promoters of the genes they transcribe. Other sequence features, often at some distance from the gene, may also participate in its transcriptional control. Most of these elements are recognised by other DNA-binding proteins generally termed 'transcription factors'. These interact in a modular fashion with the DNA control elements, and either stimulate or repress transcription. Some factors are ubiquitous, interacting with a wide range of genes in most cell types, while others are tissue-specific and may bind only to a few target genes in a single cell type. Factors binding to DNA sites at some distance from the promoter could affect transcription from that promoter in various ways. One possibility is DNA looping, such that a distantly bound transcription factor would also interact specifically with a target protein bound close to the transcriptional start site (Ptashne, 1986, 1988). In effect, the DNA loop would be cross-linked via interactions between two factors bound respectively at the proximal and distal sites. A protein bridge between two such sites is sufficient for activation even when the two are on separate DNA fragments (Muller-Storm *et al*., 1989). Moreover, many transcription factors possess separable DNA-binding and protein-binding (activator) domains.

In pol I promoters, several 'nested' control elements are located upstream from the transcriptional state site, i.e. within the spacer region separating one precursor-coding region from the next. It used to be thought that these spacers were not transcribed, based largely on EM spreads of active rDNA repeat units (see e.g. fig. 2.1B). However, more recent data show that the spacer is indeed transcribed (into very unstable RNA), both as a result of readthrough originating from the major ribosomal promoter next upstream in the cluster, and also from