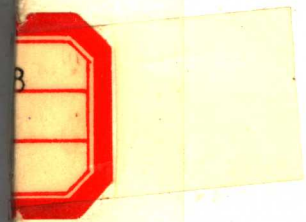
Eukaryotic Transcription Factors

David S. Latchman



Preface

In my previous book, *Gene Regulation: A Eukaryotic Perspective* (Unwin-Hyman Ltd, 1990), I described the mechanisms by which the expression of eukaryotic genes is regulated during processes as diverse as steroid treatment and embryonic development. Although some of this regulation occurs at the post-transcriptional level, it is clear that the process of gene transcription itself is the major point at which gene expression is regulated. In turn this has focused attention on the protein factors, known as transcription factors, which control both the basal processes of transcription and its regulation in response to specific stimuli or developmental processes. The characterization of many of these factors and in particular the cloning of the genes encoding them has resulted in the availability of a bewildering array of information on these factors, their mechanism of action and their relationship to each other. Despite its evident interest and importance, however, this information could be discussed only relatively briefly in *Gene Regulation*, whose primary purpose was to provide an overview of the process of gene regulation and the various mechanisms by which this is achieved.

It is the purpose of this book, therefore, to discuss in detail the available information on transcription factors, emphasizing common themes and mechanisms to which new information can be related as it becomes available. As such it is hoped the work will appeal to final-year undergraduates and postgraduate students entering the field as well as to those moving into the area from other scientific or clinical fields who wish to know how

transcription factors may regulate the gene in which they are interested.

In order to provide a basis for the discussion of transcription factors, the first two chapters focus respectively on the DNA sequences with which the factors interact and on the experimental methods which are used to study these factors and obtain the information about them provided in subsequent chapters. The remainder of the work is divided into two distinct portions. Thus Chapters 3 to 7 focus on the role of transcription factors in particular processes. These include constitutive and inducible gene expression, cell type-specific and developmentally regulated gene expression and the role of transcription factors in cancer. Chapters 8 to 10 adopt a more mechanistic approach and consider the features of transcription factors which allow them to fulfil their function. These include the ability to bind to DNA and modulate transcription either positively or negatively as well as the ability to respond to specific stimuli and thereby activate gene expression in a regulated manner.

Although this dual approach to transcription factors from both a process-oriented and mechanistic point of view may lead to some duplication, it is the most efficient means of providing the necessary overview both of the nature of transcription factors and the manner in which they achieve their role of modulating gene expression in many diverse situations.

Finally I would like to thank Mrs Rose Lang for typing the text and coping with the continual additions necessary in this fast-moving field and Mrs Jane Templeman for her outstanding skill in preparing the illustrations.

David S. Latchman

Acknowledgements

I would like to thank all the colleagues, listed below, who have given permission for material from their papers to be reproduced in this book and have provided prints suitable for reproduction.

Figure 4.3, photograph kindly provided by Dr C. Wu from Zimarino and Wu, *Nature* **327**, 727 (1987), by permission of Macmillan Magazines Ltd; Figure 4.9, photograph kindly provided by Professor M. Beato from Willmann and Beato, *Nature* **324**, 688 (1986) by permission of Macmillan Magazines Ltd. Figures 5.9 and 5.11, photographs kindly provided by Dr R. L. Davis from Davis *et al.*, *Cell* **51**, 987 (1987) by permission of Cell Press. Figures 6.1 and 10.2, photographs kindly provided by Professor W. J. Gehring from Gehring, *Science* **236**, 1245 (1987) by permission of the American Association for the Advancement of Science. Figure 6.15, photograph kindly provided by Dr P. Holland from Holland and Hogan, *Nature* **321**, 251 (1986) by permission of Macmillan Magazines Ltd. Figures 6.16 and 6.17, photographs kindly provided by Dr R. Krumlauf from Graham *et al.*, *Cell* **57**, 367 (1989) by permission of Cell Press. Figure 8.6 redrawn from Redemann *et al.*, *Nature* **332**, 90 (1988) by kind permission of Dr H. Jackle and Macmillan Magazines Ltd. Figures 8.11 and 8.15 redrawn from Schwabe *et al.*, *Nature* **348**, 458 (1990) by kind permission of Dr D. Rhodes and Macmillan Magazines Ltd; Figure 8.18 redrawn from Abel and Maniatis, *Nature* **341**, 24 (1989) by kind permission of Professor T. Maniatis and Macmillan Magazines Ltd. Figure 9.9 redrawn from Lin *et al.*, *Nature* **345**, 359 (1990) and Carey *et al.*, *Nature* **345**, 361 (1990) by kind permission of Professor M. Ptashne and Macmillan Magazines Ltd.

Contents

CHAPTER ONE

DNA sequences and transcription factors

1.1. THE IMPORTANCE OF TRANSCRIPTION

The fundamental dogma of molecular biology is that DNA produces RNA which in turn produces protein. Hence if the genetic information which each individual inherits as DNA (the genotype) is to be converted into the proteins which produce the corresponding characteristics of the individual (the phenotype), it must first be converted into an RNA product. The process of transcription whereby an RNA product is produced from the DNA is therefore an essential element in gene expression. The failure of this process to occur will obviously render redundant all the other steps which follow the production of the initial RNA transcript in eukaryotes, such as RNA splicing, transport to the cytoplasm or translation into protein (for a review of these stages see Nevins (1983)).

The central role of transcription in the process of gene expression also renders it an attractive control point for regulating the expression of genes in particular cell types or in response to a particular signal. Indeed, it is now clear that in the vast majority of cases where a particular protein is produced only in a particular tissue or in response to a particular signal this is achieved by control processes which ensure that its corresponding gene is transcribed only in that tissue or in response to such a signal (reviews: see Darnell, 1982; Latchman, 1990). For example, the genes encoding the immunoglobulin heavy and light chains of the antibody molecule are transcribed at high level only in the antibody-producing B cells (Gillis *et al.*, 1983), whilst the increase in somatostatin production in response to treatment of cells with cyclic AMP is mediated by

increased transcription of the corresponding gene (Montminy *et al.*, 1986). Therefore, while post-transcriptional regulation affecting, for example, RNA splicing or stability plays some role in the regulation of gene expression (reviews: Brawerman, 1987; Breitbart *et al.*, 1987), the major control point lies at the level of transcription.

1:2 DNA SEQUENCE ELEMENTS

1.2.1 The gene promoter

The central role of transcription both in the basic process of gene expression in and its regulation in particular tissues has led to considerable study of this process. Initially such studies focused on the nature of the DNA sequences within individual genes which were essential for either basal or regulated gene expression. In prokaryotes, such sequences are found immediately upstream of the start site of transcription and form part of the promoter directing expression of the genes. Sequences found at this position include both elements found in all genes which are involved in the basic process of transcription itself and those found in a more limited number of genes which mediate their response to a particular signal (Schmitz and Galas, 1979; Miller and Reznikoff, 1980; Ptashne, 1986).

Early studies of cloned eukaryotic genes therefore concentrated on the region immediately upstream of the transcribed region where, by analogy, sequences involved in transcription and its regulation should be located. Putative regulatory sequences were identified by comparison between different genes, and the conclusions reached in this way confirmed either by destroying these sequences by deletion or mutation or by transferring them to another gene in an attempt to alter its pattern of regulation.

This work, carried out on a number of different genes encoding specific proteins, identified many short-sequence elements involved in transcriptional control (reviews: Davidson *et al.*, 1983; Jones *et al.*, 1988). The elements of this type present in two typical examples, the human gene encoding the 70-kDa heat-inducible (heat-shock) protein (Williams *et al.*, 1989) and the human metallothionein IIA gene (Lee *et al.*, 1987a), are illustrated in Figure 1.1.

Comparisons of these and many other genes revealed that, as in bacteria, their upstream regions contain two types of elements: firstly, sequences found in very many genes exhibiting distinct patterns of regulation which are likely to be involved in the basic

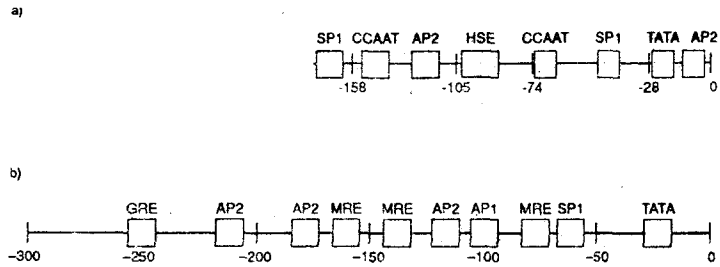


Figure 1.1 Transcriptional control elements upstream of the transcriptional start site in the human genes encoding hsp70 (a) and methallothionein IIA (b). The TATA, Sp1 and CCAAT boxes bind factors which are involved in constitutive transcription (see also Chapter 3) whilst the glucocorticoid response element (GRE), metal response element (MRE), heat shock element (HSE) and the AP1 and AP2 sites bind factors involved in the induction of gene expression in response to specific stimuli (see also Chapter 4 and Section 7.2).

process of transcription itself; and secondly, those found only in genes transcribed in a particular tissue or in response to a specific signal which are likely to produce this specific pattern of expression. These will be discussed in turn.

1.2.2 Sequences involved in the basic process of transcription.

Although they are regulated very differently, the *hsp70* and metallothionein genes both contain a TATA box. This is an AT-rich sequence (consensus TATAA/TAA/T) which is found about 30 base pairs upstream of the transcriptional start site in very many but not all genes. Mutagenesis or relocation of this sequence has shown that it plays an essential role in accurately positioning the start site of transcription (Breathnach and Chambon, 1981). The region of the gene bracketed by the TATA box and the site of transcriptional initiation (the Cap site) has been operationally defined as the gene promoter (Goodwin *et al.*, 1990). It is likely that this region binds several proteins essential for transcription, as well as RNA polymerase II itself, which is the enzyme responsible for transcribing protein-coding genes (Lewis and Burgess, 1982).

Although the TATA box is found in most eukaryotic genes, it is absent in some genes, notably housekeeping genes expressed in all tissues and in some tissue-specific genes (reviews: Sehgal *et al.*,

1988; Smale and Baltimore, 1989). In these promoters, the actual sequence over the start site of transcription itself appears to play a critical role in determining the initiation point and acts as a minimal promoter capable of producing basal levels of transcription (Smale and Baltimore, 1989).

In promoters which contain a TATA box and in those which lack it, the very low activity of the promoter itself is dramatically increased by other elements located upstream of the promoter. These elements are found in a very wide variety of genes with different patterns of expression, indicating that they play a role in stimulating the constitutive activity of promoters. Thus inspection of the *hsp70* and metallothionein IIA genes reveals that both contain one or more copies of a GC-rich sequence known as the Sp1 box which is found upstream of the promoter in many genes both with and without TATA boxes (review: Dynan and Tjian, 1985).

In addition the *hsp70* promoter but not the metallothionein promoter contains another sequence, the CCAAT box, which is also found in very many genes with disparate patterns of regulation. Both the CCAAT box and the Sp1 box are typically found upstream of the TATA box, as in the metallothionein and *hsp70* genes. Some genes, as in the case of hsp70, may have both of these elements, whereas others, such as the metallothionein gene, have single or multiple copies of one or the other (review: McKnight and Tjian, 1986). In every case, however, these elements are essential for transcription of the genes, and their elimination by deletion or mutation abolishes transcription (McKnight and Kingsbury, 1982; McKnight *et al.*, 1984). Hence these sequences play an essential role in efficient transcription of the gene and have been termed upstream promoter elements (UPE) (Goodwin *et al.*, 1990). The role of the promoter and UPE sequences and the protein factors which bind to them are discussed further in Chapter 3.

1.2.3. Sequences involved in regulated transcription

Inspection of the *hsp70* promoter (Figure 1.1) reveals several other sequence elements which are only shared with a much more limited number of other genes and which are interdigitated with the upstream promoter elements discussed above. Indeed, one of these, which is located approximately 90 bases upstream of the transcriptional start site, is shared only with other heat-shock genes whose transcription is increased in response to elevated temperature. This suggests that this heat-shock consensus element may be essential for the regulated transcription of the *hsp70* gene in response to heat.

To directly prove this, however, it is necessary to transfer this sequence to a non-heat-inducible gene and show that this transfer renders the recipient gene heat-inducible. Pelham (1982) successfully achieved this by linking the heat-shock consensus element to the non-heat-inducible thymidine kinase gene of the eukaryotic virus herpes simplex. This hybrid gene could be activated following its introduction into mammalian cells by raising the temperature (Figure 1.2). Hence the heat-shock consensus element can confer heat inducibility on another gene, directly proving that its presence in the *hsp* gene promoters is responsible for their heat inducibility.

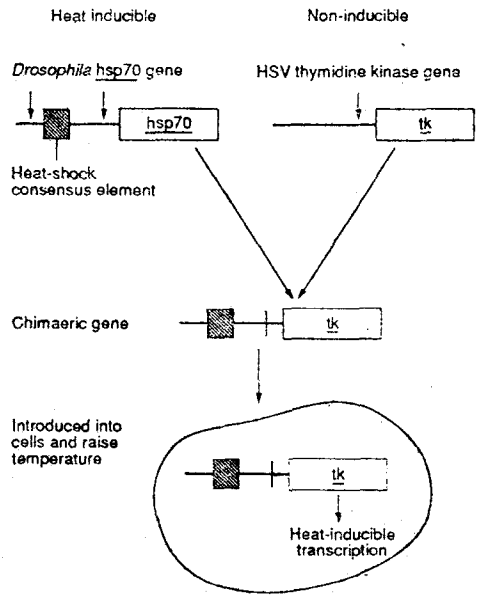


Figure 1.2 Demonstration that the heat-shock consensus element mediates heat inducibility. Transfer of this sequence to a gene (thymidine kinase) which is not normally inducible renders this gene heat-inducible.

Moreover, although these experiments used a heat-shock consensus element taken from the *hsp70* gene of the fruit fly *Drosophila melanogaster*, the hybrid gene was introduced into mammalian cells. Not only does the successful functioning of the fly element in mammalian cells indicate that this process is evolutionarily conserved but it permits a further conclusion about the way in which the effect operates. Thus in the cold-blooded *Drosophila*, 37°C

represents a thermally stressful temperature and the heat-shock response would normally be active at this temperature. The hybrid gene was inactive at 37°C in the mammalian cells, however, and was only induced at 42°C, the heat-shock temperature characteristic of the cell into which it was introduced. Hence this sequence does not act as a thermostat, set to go off at a particular temperature, since this would occur at the *Drosophila* heat-shock temperature (Figure 1.3a). Rather, this sequence must act by being recognized by a cellular protein which is activated only at an elevated temperature characteristic of the mammalian cell heat-shock response (Figure 1.3b).

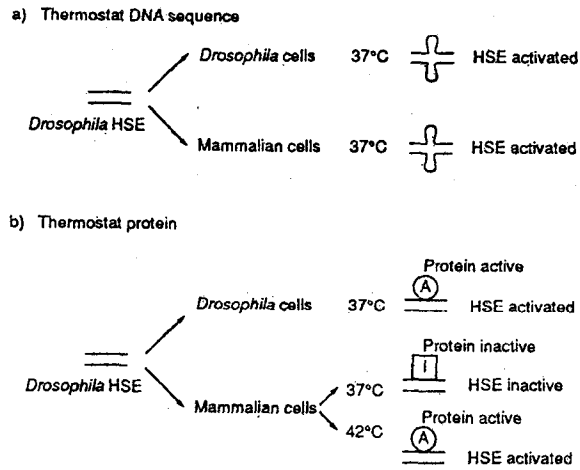


Figure 1.3 Predicted effects of placing the *Drosophila* heat-shock element in a mammalian cell if the element acts as a thermostat detecting elevated temperature directly (a) or if it acts by binding a protein which is activated by elevated temperature (b). Note that only possibility (b) can account for the observation that the *Drosophila* heat-shock element only activates transcription in mammalian cells at the mammalian heat-shock temperature of 42°C and not at the *Drosophila* heat-shock temperature of 37°C.

This experiment therefore not only directly proves the importance of the heat-shock response element in producing the heat inducibility of the *hsp70* gene but also shows that this sequence acts by binding a cellular protein which is activated in response to elevated temperature. The binding of this transcription factor then activates transcription of the *hsp70* gene. The manner in which this factor activates transcription of the *hsp70* gene and the other heat-shock genes is discussed further in Section 4.2.

The presence of specific DNA sequences which can bind particular proteins will therefore confer on a specific gene the ability to respond to particular stimuli. Thus the lack of a heat-shock consensus element in the metallothionein IIA gene (Figure 1.1) means that this gene is not heat-inducible. In contrast, however, this gene, unlike the *hsp70* gene, contains a glucocorticoid response element (GRE). Hence it can bind the complex of the glucocorticoid receptor and the hormone itself which forms following treatment of cells with glucocorticoid (review: Beato, 1989). Its transcription is therefore activated in response to glucocorticoid whereas that of the *hsp70* gene is not (see Section 4.3). Similarly, only the metallothionein gene contains metal response elements (MRE), allowing it to be activated in response to treatment with heavy metals such as zinc and cadmium (Karin *et al.*, 1984). In contrast, both genes contain binding sites for the transcription factor AP2 which mediates gene activation in response to cyclic AMP and phorbol esters (Imagawa *et al.*, 1987). The manner in which the binding of specific transcription factors to different regulatory sequences modulates gene expression in response to specific inducing factors is discussed further in Chapter 4.

Similar DNA sequence elements in the promoters of tissue-specific genes play a critical role in producing their tissue-specific pattern of expression by binding transcription factors which are present in an active form only in a particular tissue where the gene will be activated. For example, the promoters of the immunoglobulin heavy- and light-chain genes contain a sequence known as the octamer motif, ATGCAAAT (Parslow *et al.*, 1984; Mason *et al.*, 1985) which can confer B cell-specific expression on an unrelated promoter (Wirth *et al.*, 1987). This sequence acts by binding a transcription factor known as Oct-2 which is absent in most cell types but is expressed at high levels in immunoglobulin-producing B cells (Staudt *et al.*, 1986; Landolfi *et al.*, 1986) (see Section 5.2.2 for further details). Similarly, the related sequence ATGAATAA/T is found in two genes expressed specifically in the anterior pituitary gland, the prolactin gene and the growth hormone gene (Nelson *et al.*, 1988), and binds a transcription factor known as Pit-1, which is expressed only in the anterior pituitary (Ingraham *et al.*, 1988). If this short sequence is inserted upstream of a promoter, the gene is expressed only in pituitary cells. In contrast, the octamer motif, which differs by only two bases, will direct expression only in B cells when inserted upstream of the same promoter (Elsholtz *et al.*, 1990; Figure 1.4). Hence small differences in control-element sequences can produce radically different patterns of gene expression.

The role of transcription factors in producing tissue specific gene expression is discussed further in Chapter 5.

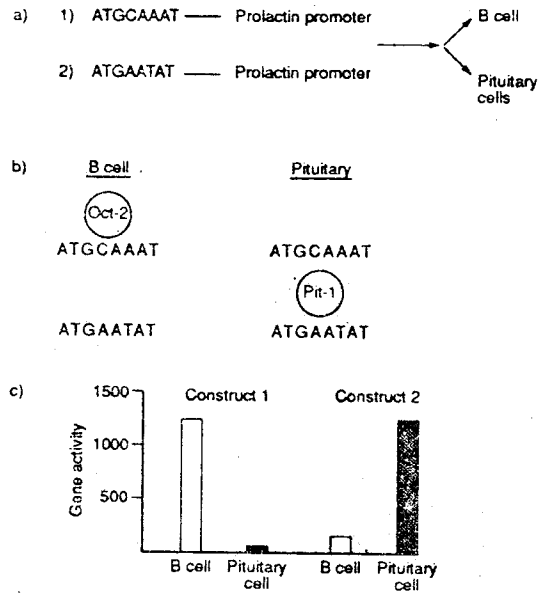


Figure 1.4 Linkage of the octamer-binding motif ATGCAAAT (1) and the related Pit-1-binding motif ATGAATAT (2) to the prolactin promoter and introduction into B cells and pituitary cells (a). In B cells, the octamer motif binds the B-cell-specific octamer binding protein Oct-2, whereas the Pit-1 motif fails to bind any protein. In contrast, in pituitary cells the Pit-1 motif binds the Pit-1 protein whereas the octamer motif fails to bind any protein (b). This results in only the octamer containing construct 1 directing a high level of activity in B cells, whereas only construct 2 containing the Pit-1-binding site directs a high level of gene activity in pituitary cells (c). Data from Elsholtz *et al.*, (1990).

1.2.4 Enhancers

One of the characteristic features of eukaryotic gene expression is the existence of sequence elements located at great distances from the start site of transcription which can influence the level of gene expression. These elements can be located upstream, downstream or within a transcription unit and function in either orientation relative to the start site of transcription (Figure 1.5). They act by

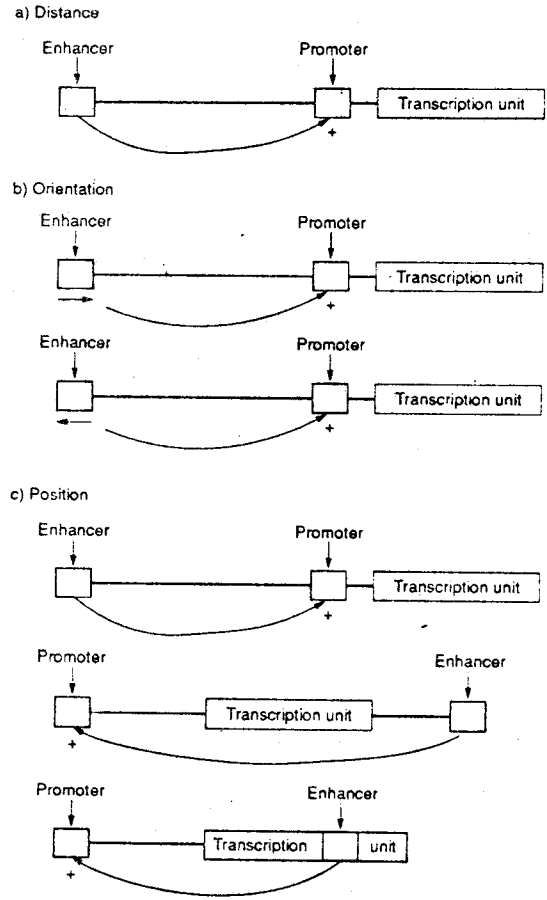


Figure 1.5 Characteristics of an enhancer element which can activate a promoter at a distance (a); in either orientation relative to the promoter (b); and when positioned upstream, downstream, or within a transcription unit (c).

increasing the activity of a promoter, although they lack promoter activity themselves and are hence referred to as enhancers (reviews: Serfling *et al.*, 1985; Hatzopoulos *et al.*, 1988). Some enhancers are active in all tissues and increase the activity of a promoter in all cell types whilst others function as tissue-specific enhancers which activate a particular promoter only in a specific cell type. Thus the enhancer located in the intervening region of the immunoglobulin

genes is active only in B cells (Gillis *et al.*, 1983) and the B cell-specific expression of the immunoglobulin genes is produced by the interaction of this enhancer and the immunoglobulin promoter which, as we have previously seen, is also B cell-specific (Garcia *et al.*, 1986) (see Section 5.2 for further discussion).

As with promoter elements, enhancers contain multiple binding sites for transcription factors which interact together (review: Dynan, 1989). In many cases these elements are identical to those contained immediately upstream of gene promoters. Thus the immunoglobulin-heavy chain enhancer contains a copy of the octamer sequence (Sen and Baltimore, 1986) which is also found in the immunoglobulin promoters (Section 1.2.3). Similarly, multiple copies of the heat-shock consensus element are located far upstream of the start site in the *Xenopus hsp70* gene and function as a heat-inducible enhancer when transferred to another gene (Bienz and Pelham, 1986).

Enhancers therefore consist of sequence elements which are also present in similarly regulated promoters and may be found within the enhancer associated with other control elements or in multiple copies.

1.3 INTERACTION BETWEEN FACTORS BOUND AT VARIOUS SITES

The typical eukaryotic gene will therefore consist of up to four distinct transcriptional control elements (Figure 1.6). These are:

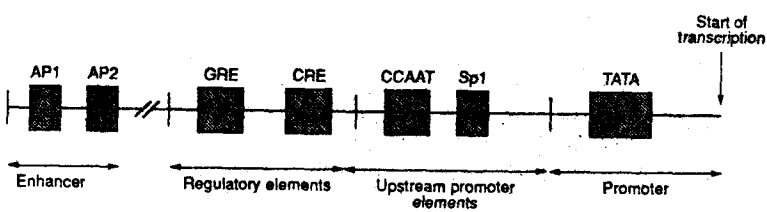


Figure 1.6 Structure of a typical gene with a TATA box-containing promoter, upstream promoter elements such as the CCAAT and Sp1 boxes, regulatory elements inducing expression in response to treatment with substances such as glucocorticoid (GRE) and cyclic AMP (CRE) and other elements within more distant enhancers. Note that, as discussed in the text and illustrated in Figure 1.1, the upstream promoter elements are often interdigitated with the regulatory elements whilst the same regulatory elements can be found upstream of the promoter and in enhancers.