J. Biosci., Vol. 18, Number 1, March 1993, pp 27-36. © Printed in India.

Specificity in the regulation of eukaryotic gene transcription*

G PADMANABAN

Department of Biochemistry, Centre for Genetic Engineering, and Jawaharlal Nehru Centre for Advanced Scientific Research, Indian Institute of Science, Bangalore 560 012, India

MS received 13 July 1992; revised 30 October 1992

Abstract. The regulation of eukaryotic gene transcription poses major challenges in terms of the innumerable protein factors required to ensure tissue or cell-type specificity. While this specificity is sought to be explained by the interaction of *cis*-acting DNA elements and the *trans*-acting protein factor(s), considerable amount of degeneracy has been observed in this interaction. Immunoglobulin heavy chain gene expression in B cells and liver-specific gene expression are discussed as examples of this complexity in this article. Heterodimerization and post-translational modification of transcription factors and the organization of composite promoter elements are strategies by which diverse sets of genes can be regulated in a specific manner using a finite number of protein factors.

Keywords. Transcription in eukaryotes; regulation.

1. Introduction

Different mammalian cell types have essentially the same DNA complement and do not differ in terms of sequence potential for encoding a variety of proteins. Nevertheless, production of haemoglobin is restricted to erythroid cells, insulin to β cells in the islets of Langerhans of the pancreas, albumin to liver cells, and ovalbumin to oviduct cells. While powerful techniques may be able to detect traces of these proteins or their mRNAs in other tissue cells as well, it is clear that these genes show tissue and cell-type specificity for optimal expression. How do cells regulate the genetic machinery for synthesis of the ubiquitous housekeeping proteins, but impart specificity in the expression of other proteins characterizing the cell type or its differentiated state?

While the many steps involved in the transfer of information from the gene to the protein can all be subject to regulation, transcription is a major step where specificity of gene expression is regulated. Recent studies have laid a framework which explains how this specificity can be achieved with respect to transcription mediated by RNA polymerase II. This forms the subject matter of this article.

2. *Cis*-acting DNA elements and *trans* -acting protein factors

A scheme for the formation of the initiation complex in transcription has been described using the adenovirus major late (AdML) promoter as the prototype. The

*Based on the lecture given at the Symposium on "Regulation of Gene Expression" held in Bangalore on January 20-21, 1992.

general transcription factor TFIID binds to the TATA element located at –31 to –25 from the transcription initiation site. TFIIA stabilizes the TFIID-DNA complex and then TFIIB binds to this complex, providing a scaffold for the binding of RNA polymerase II. This binding takes place in association with TFIIF, followed by the binding of TFIIE. At this stage, the promoter is converted to the open state, providing the appropriate condition for initiation of RNA synthesis (Parvin *et al* 1992).

Regulation of transcription is achieved by the interaction of *cis*-acting DNA elements and *trans*-acting protein factors (Wasylyk 1988; Johnson and McKnight 1989), Promoters, enhancers, and their negative counterparts (silencers) constitute *cis*-acting elements, which are found in the 5' or 3' flanking regions of the gene, and within introns. The overall architecture of *cis*-acting elements — their multiplicity, degeneracy and distance from the transcription start site—decides the response of the cell in terms of expressing a particular gene. This response is achieved through the interaction of a variety of protein factors, some of which ultimately make contact with the general transcription factors and RNA polymerase II to activate or repress transcription (Sawadogo and Sentenac 1990). Long-distance protein–protein contacts are facilitated by DNA bending (Ptashne 1988).

Although the step of transcription initiation has been looked at as the most probable site of regulation (Wasylyk 1988), recent studies do indicate that transcript elongation is another step for regulation (Spencer *et al* 1990). Very little is known about the features of regulation of transcription at the level of termination. These mechanisms of control of gene expression do not negate the overall importance of chromatin architecture in transcriptional regulation. The role of nucleosome structure, and the importance of histones in particular, which were prominently investigated some years ago, got relegated to the background because of the popularity of the subject of *cis*-acting DNA elements and specific *trans*-acting protein factors. However, the concept that chromatin architecture may be involved, not only as a global regulatory format but also in the regulation of specific genes, is slowly gaining ground. This feature operates at the global level of demarcating the euchromatic and heterochromatid regions, and at the specific gene level by influencing (i) structure at the DNA site for interaction with specific protein factors and (ii) the mobility of the transcription machinery on the DNA template (Roeder 1991).

If it is granted that the specificity in the regulation of gene transcription basically rests in the specificity of interaction of *cis*-acting elements with *trans*-acting factors, then the question arises as to how many protein factors are needed to produce one protein. If this argument is extended, the incongruous situation of the entire genetic potential of eukaryotic DNA not being adequate to regulate transcription would arise. With the knowledge that the bulk eukaryotic DNA remains silent at any one time, it is obvious that a far more efficient machinery to regulate specificity in gene transcription is operative. Recent studies indicate that mechanisms have evolved that permit degeneracy in DNA-protein interaction, at the same time ensuring specificity through auxiliary DNA-protein and protein-protein contacts, such that a finite number of *trans*-acting factors can differentially interact with different architectures of *cis*-acting elements. This can be illustrated with a few examples.

3. B-cell specificity in immunoglobulin heavy chain gene expression (Schreiber *et al* 1989)

During B-cell development there is rearrangement of the immunoglobulin heavy (IgH) chain gene, when one of the several hundred V (variable) region genes recombines with D (diversity) and J (joining) segments. This VDJ segment is separated from the C (constant) region by a large intron. A strong enhancer is located in the intron which becomes active only after the recombination (figure 1). Although the germ-line V_H gene contains a promoter, it is only basally active. After the rearrangement, one of the V region promoters is brought under the control of the enhancer, leading to strong transcription of functional IgH mRNA starting at the VDJ cap site.

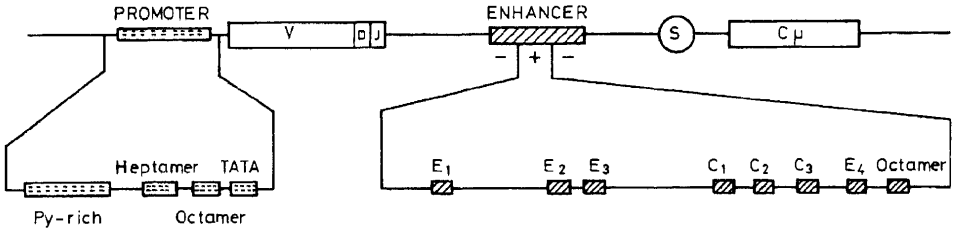

Figure 1. Architecture of the regulatory elements in the mouse immunoglobulin heavy chain gene. The figure is based on the figure in Schreiber *et al* (1989).

3.1 *Structure of the mouse IgH enhancer*

The IgH enhancer has proven to be highly B-cell-specific and is not functional when transfected into fibroblasts or epithelial cells. The enhancer activity has been narrowed down to a 224 bp HinfI fragment and consists of El, E2, E3, E4, Cl, C2, C3 and octamer motifs (figure 1). Each one of these motifs contributes to the total activity of the enhancer, the octamer being the strongest motif, since mutation at the octamer site brings down transcription by 50%. Multimerization of the octamer motif within a 51-bp DdeI-HinfI fragment yields substantial B-cell-specific enhancer activity.

3.2 *Structure of V_H promoter*

This consists of the TATA box, the octamer sequence located at – 60 to – 70 bp, a heptamer sequence adjacent to the octamer sequence, and a pyrimidine-rich sequence further upstream (figure 1). The octamer sequence is the most crucial for activity of the V_H promter in B lymphoid cells. An interesting feature is that even though the octamer and heptamer sequences are quite different the same factor binds to both the sequences. Nevertheless, the octamer motif alone is sufficient to create a lymphoid-cell-specific promoter when placed upstream of a heterologous TATA box.

The crucial role of the octamer sequence in B-cell-specific expression is thus indicated by the facts that (i) it can constitute a B-cell-specific promoter when

present along with the canonical TATA box and (ii) multiple octamer motifs can create a B-cell-specific enhancer.

3.3 *The octamer paradox*

Although the octamer motif confers B-cell specificity, it is widely present in housekeeping genes. Even if stringent criteria, such as presence of certain adjacent nucleotides, are applied, the motif is still detected in, for example, histone H2B promoter of sea urchin, adenovirus and a U2snRNA gene of *Xenopus laevis.* A ubiquitous octamer-binding protein, OTFl, has also been detected in all these cell types.

Subsequently, however, a B-cell-specific octamer-binding protein OTF2A and yet another, OTF2B, have been identified. The most convincing proof that OTF2A is B-cell-specific comes from transfection experiments with HeLa cells using an expression vector for OTF2A and a reporter gene controlled by B-cell-specific promoter and SV40 enhancer. A strong stimulatory effect on reporter-gene expression, but not in any of the appropriate controls, is clear proof for the change of cell-type specificity by OTF2A. The question arises as to why OTFl, which is ubiquitously present and interacting with the octamer motif, fails to confer B-cell specificity. Models can be proposed to explain this situation, and one scenario can be that while OTF2A is functional by itself OTFl can be functional only in conjunction with adjacent binding of the Spl factor or the CCAAT-binding protein. In addition to OTF2A not being expressed in non-B cells, negatively acting sequences in the IgH enhancer appear to prevent accessibility of the Ig gene locus to OTFl in non-B cells. The absence of DNA rearrangement in non-B cells would render the distance between IgH promoter and enhancer too large for productive interaction. In B cells, it is conceivable that OTF1 as well as OTF2A/OTF2B also participate in housekeeping functions.

Thus, despite the apparent lack of specificity of the octamer motif in its interaction with OTFl, OTF2A and OTF2B proteins and the ubiquitous presence of the motif, cell-type specificity can be achieved by invoking auxiliary DNA-protein interactions with positive and negative modulatory effects as well as a synergistic interaction between the octamer motifs in the enhancer and promoter regions through protein-protein interaction (Gerster *et al* 1987; Wirth *et al* 1987).

4. Specificity in liver gene expression (Lai and Darnell 1991)

In recent years, typical genes expressed in liver, such as albumin, α-fetoprotein, α-1-antitrypsin, fibrinogen, transthyretin and a few others, have come in for detailed investigation from the perspective of tissue-specific or cell-specific gene expression as well as the possible role some of the liver-specific transcription factors play in the making of developmental decisions which affect endodermal cells.

The transcription factors that confer liver-specific gene expression have been named C/EBP, HNF1, HNF3 and HNF4 families. Most of these factors react with multiple promoter sites at near-5' upstream of the liver genes mentioned and appear to drive their expression. Of these, the consensus sequence for HNF 1 binding has been identified in the liver genes mentioned. However, the binding sites for C/EBP

are pleomorphic and there is no recognized consensus DNA binding site for HNF4. Thus, binding studies *in vitro* and transfection assays in cell lines do not fully reveal the specificities generated *in vivo.* For example, *in vivo* footprint analysis on the transthyretin gene in adult liver indicates the involvement of only a subset of sites detected by transfection experiments in HeLa cells. HNF3 sites are well protected against DNase I or chemical attack, but HNF4 and C/EBP sites are not protected. Therefore, if the latter sites are involved *in vivo,* they may, perhaps, participate in establishing active transcription sites but may not be required thereafter.

The question of tissue specificity can be illustrated with the following example: HNF1 and HNF4 are present in the liver and kidney, but the genes that have binding sites for these factors in the kidney are not active or at least not as active as in the liver. This situation suggests that negative regulation could be one mechanism operating. In transgenic animals, deletion of certain regulatory sites has been shown to cause inappropriate expression of the α-fetoprotein gene. In the retinol binding protein gene, removal of an 80-bp fragment results in its expression in HeLa cells as well as hepatoma cells. Among the HNF1 family, HNF1β is very similar to HNFlα in the DNA-binding and dimerization domains, but not in the activation domain. Thus, HNF1 β expression is correlated with the repression of a subset of liver genes. Another interesting feature of liver transcription factors is their relation to homeotic genes in *Drosophila.* HNF1, through its DNA-binding region, is distantly related to homeobox proteins. In HNF3, 86 out of 110 amino acids in the DNA-binding region are identical with those in the product of *forkhead,* a *Drosophila* homeotic gene. HNF4 has high homology in its DNA-binding, ligand-binding or dimerization domains to the *Drosophila* homologue than to any mammalian nuclear receptor. There is evidence for the presence of these liver transcription factors in rat embryos at mid-gestation. These factors may therefore play a role in endodermal differentiation decisions in the primitive gut in a fashion analogous to the effect of homeobox proteins on anterior-posterior axis patterning in *Drosophila.*

Studies in my laboratory have shown that the CYP2B1/B2 family of the cytochrome P-450 gene superfamily, induced in rat liver by the prototype drug phenobarbitone, may set a new paradigm for tissue specificity of gene expression. First of all, the genes do not seem to have the consensus *cis*-acting elements in the near-5' upstream region observed with some of the other liver-specific genes (Rangarajan *et al* 1987). Secondly, although detailed sequence information is not yet available, the positive *cis*-acting element identified in the near-5' upstream region of the CYP2B1/B2 genes (Rangarajan and Padmanaban 1989; Poornima *et al* 1992) may not occur upstream of other phenobarbitone-inducible genes. Thirdly, there are chemicals such as allylisopropylacetamide, which bear apparently no structural similarity to phenobarbitone, but nevertheless activate CYP2B1/B2 gene transcription (Dwarki *et al* 1987). He and Fulco (1991) have observed that a 17-bp fragment within the positive *cis*-acting element identified in my laboratory is present in the CYP2B1/B2 genes of rat liver and the barbiturate-inducible gene in *Bacillus megaterium.* However, while the 17-bp fragment may act as a positive element in the rat, it appears to function as a negative element in the bacterium. Although further details of this regulation are awaited, studies in my laboratory reveal that phosphorylation status of the transcription factors may be the crucial property affected by the inducer (unpublished data). It is thus clear that there is yet

another dimension to the phenomenon of tissue specificity in the regulation of gene transcription. Specificity can be imparted by post-translational modification of a transcription factor, the binding or lack of binding of which to the same core sequence may result in opposite effects in different systems. The sequence flanking the core sequence and auxiliary protein factors should obviously play a role in deciding the specificity.

5. Mechanisms to achieve diversity and specificity in transcriptional regulation

5.1 *Interaction of transcription factors with DNA as dimers* (*Lamb and McKnight 1991)*

Although the ability to dimerize is not a universal phenomenon among transcription factors, homodimerization and heterodimerization are frequently observed. Several classes of dimerization interface have been identified. The leucine zipper motif is the most extensively studied. First detected in C/EBP, many proteins such as Jun, Fos and Myc have been found to contain a heptad repeat of leucines within a stretch of 30 amino acids, capable of forming an α-helix. The 'zippering' of two such helices, facilitated by the attractive interactions between the leucine residues partitioned along the hydrophobic face of each helix, facilitates dimerization. The basic regions adjacent to the zipper region make contacts with the DNA (figure 2).

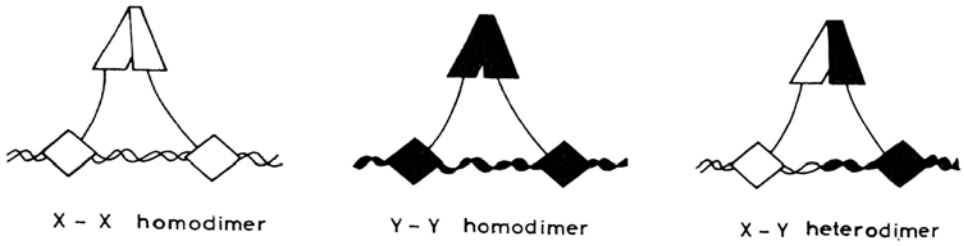

Figure 2. Alteration of DNA-binding specificity by heterodimerization. The heterodimeric zipper formation depicted at right permits binding to a hybrid DNA-recognition site (unshaded and shaded DNA and basic-residue-binding regions). The figure is based on the figure in Lamb and McKnight (1991).

Transcription factors can be grouped into subfamilies based on their ability to form heterodimers, and dimerization among the members is not promiscuous. For example, members of the C/EBP family will heterodimerize with each other, but not with those of Fos, Jun or ATF-CREB (activating transcription factor; cAMP response element binding protein). Many eukaryotic proteins bind to sites on DNA that show dyad symmetry, each monomer of the dimer contacting one half of the dyad substrate (figure 2).

A less-well-known motif of dimerization is exemplified by helix loop helix (HLH) proteins. The protein Id, an HLH protein, lacks the basic amino acids to bind to DNA, but can heterodimerize with MyoD (in erythroid cells) and prevent its

function. Similarly, in *Drosophila*, the Emc and hairy gene products can, through heterodimerization, interfere with the positive influence of AS-C and Daughterless on the development of cells competent to form sensory organs.

These dimerization motifs confer not only DNA-binding specificity, but also specificity in making contacts with other proteins. For example, the members of the C/EBP family (α, *β* and δ), although located on different chromosomes, not only show a remarkable conservation of amino acids in the DNA-binding face of the α-helical region that gets embedded in the major groove, but the sequence of the solvent-exposed face is also highly conserved. This high degree of conservation of the solvent-exposed face may reflect its accessibility to protein-protein contacts critical to the function of all members of the C/EBP family. Thus, heterodimerization and specificity in establishment of distant protein contacts can account for the diversity generated from a limited number of transcription factors, while at the same time ensuring specificity.

5.2 *Crosstalk at composite response elements* (*Miner and Yamamoto 1991*)

This model allows for cell-specific effects by ubiquitous factors and divergent effects by closely related factors from a given family (figure 3).

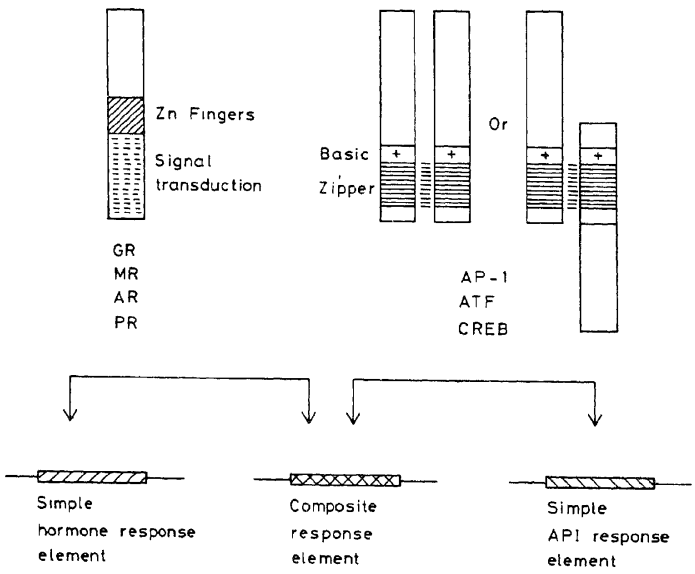

Figure 3. Specificity of composite DNA elements. The bars at the top represent two families of transcription factors. The one on the left represents steroid receptors with zinc finger motifs. The ones on the right represent the AP1 family whose members have the leucine zipper motif. The composite element is capable of interacting with members of both families. The figure is based on the figure in Miner and Yamamoto (1991). GR, Glucocorticoid receptor; MR, mineralocorticoid receptor; PR, progesterone receptor; AR, androgen receptor; AP1, activator protein 1; CREB, cAMP response element binding protein; ATF, activating transcription factor.

For example, the glucocorticoid receptor (GR) acts through a zinc finger DNA-binding domain in presence of the hormone, assuming a dimer conformation. In the absence of DNA the receptor is predominantly monomeric. On the other hand, the AP1 factors (Fos and Jun subfamilies) form a dimer even in the absence of DNA through a leucine zipper, while the basic region of the proteins contacts the tissue plasminogen activator (TPA) response element (TRE). It has been shown that the proliferin promoter is regulated by TPA and glucocorticoid, and an oligonucleotide corresponding to the bound sequence, denoted pIfG, binds purified GR as well as AP1. This composite element does not appear to be related to the simple glucocorticoid response element (GRE), and shows remarkable cell specificity. Such specificity does not require special factors, but is based on the ratio of the ubiquitous factors c-Jun and c-Fos. For example, F9 cells, which lack AP1 activity under appropriate culture conditions, displayed no hormonal regulation of pIFG-linked reporter CAT gene. HeLa cells, which express AP1 predominantly as c-Jun homodimers, enhanced reporter expression in response to dexamethasone, CV-1 cells, in which AP1 is comprised mostly of c-Fos/c-Jun heterodimers, repress pIfG promoter expression in response to the hormone. These regulatory connections involve physical interaction between GR and AP1.

The second feature, namely of divergent effects by closely related factors from a given family, is exemplified by observations with Jun- and Fos-related families. Jun and the related members in the ATF/CREB subfamily as well as Fos and its related members in the ATF/CREB subfamily behave similar to Jun or Fos in terms of dimerization or DNA-binding specificity. It is not clear why there should be many members with similar specificities. However, the individual members within a subfamily show differences at composite elements. For example, c-Fos and Fra- 1 from the Fos subfamily, and GR and MR from the receptor family show distinctly different or even opposite behaviour at composite elements.

More than one model can be proposed to explain interactions at the composite regulators. A 'co-occupancy' scheme, where factors from different regulatory families interact with DNA as well as with each other, has been proposed. Factor tethering refers to two different proteins establishing protein-protein contacts on the DNA, but only of the two actually binding to the DNA. In the third mechanism, two factors from different families interact with each other in absence of DNA, but each blocks the other from binding to DNA. Specific examples are available that favour one or the other mechanism.

6. Conclusions

It is obvious that eukaryotes possess complex mechanisms to achieve specificity in the regulation of gene transcription, unlike prokaryotes, where repressor/activator interactions with promoter elements are simpler. The reason for this situation in eukaryotes is perhaps due to the necessity to achieve regualtion of specific gene transcription while at the same time ensuring diversity of a functional architecture generated from a finite number of transcription factors. Thus, dimerization phenomena and post-translational modifications at the protein level and composite promoter element architecture at the DNA level offer strategies to achieve this goal. In fact, this versatility may not be restricted to gene-specific or cell-specific protein

factors (Green 1992). The belief that the general transcription factors (TFIIA, TFIIB, TFIIC, TFIID and TFIIE) may participate uniformly in all Pol II-mediated transcription is undergoing a change. Each general transcription factor need not perform a unique function. More than one protein fraction appears to have TFIIA activity. TFIIE is required for AdML promoter-driven transcription, but does not appear to be required for IgH promoter-driven transcription. Although possibilities such as a different protein functioning as TFIIE with IgH promoter or requirement of very low amounts of TFIIE for IgH transcription have not been ruled out, it is clear that degeneracy of function will be seen at all levels. It appears that TATA-binding proteins may function not only in Pol II-mediated transcription, but also in Pol I-mediated and Pol III-mediated transcription. The fascinating pursuit of the solution of the jigsaw puzzle of regulation of transcription of a large number of genes under a wide variety of conditions in a specific manner with a finite number of genetically coded factors will pose a challenge for years to come.

Acknowledgements

Research from the author's laboratory was supported by grants from the Department of Science and Technology, New Delhi. GP holds the C V Raman Professorship of the Indian National Science Academy, New Delhi.

References

Dwarki V J, Francis V SN K, Bhat G J and Padmanaban G.1987 Regulation of cytochrome P-450 gene transcription, messenger RNA and apoprotein levels by heme; *J. Biol. Chem.* **262** 16958–16962

Gerster T, Matthias P, Thali M, Jiricny J and Schafner W 1987 Cell type specificity elements of the immunoglobulin heavy chain gene enhancer; *EM BO J.* **6** 1323–1330

Green M R 1992 Transcriptional transgressions; *Nature* (*London)* **357** 364–365

He J-S and Fulco A J 1991 A barbiturate-regulated protein binding to a common sequence in the cytochrome P-450 genes of rodents and bacteria; *J. Biol. Chem.* **266** 7864–7869

Johnson P F and McKnight S L 1989 Eukaryotic transcriptional regulatory proteins; *Annu. Rev. Biochem.* **58** 799–839

Lai E and Darnell J E Jr 1991 Transcriptional control in hepatocytes; *Trends Biochem. Sci.* **16** 427-430

Lamb P and McKnight S L 1991 Diversity and specificity in transcriptional regulation: the benefits of heterotypic dimerization; *Trends Biochem. Sci.* **16** 417–422

Miner J N and Yamamoto K 1991 Regulatory crosstalk at composite response elements; *Trends Biochem. Sci.* **16** 423–426

Parvin J D, Marc Timmers H Th and Sharp P A 1992 Promoter specificity of basal transcription factors; *Cell* **68** 1135–1144

Ptashne M 1988 How eukaryotic transcriptional activators work; *Nature* **335** 683–689

Poornima U, Venkateswara Rao M, Venkateswar V, Rangarajan P N and Padmanaban G 1992 Identification and functional characterization of a cis-acting positive DNA element regulating CYP2B1/B2 gene transcription in rat liver; *Nucleic Acids Res.* **20** 557–562

Rangarajan P N, Ravishankar H and Padmanaban G 1987 Isolation of a cytochrome P-450e gene variant and characterization of its 5'-flanking sequences; *Biochem. Biophys. Res. Commun.* **114** 258–263

Rangarajan P N and Padmanaban G 1989 Regulation of cytochrome P-450 b/e gene expression by a heme- and phenobarbitone-modulated transcription factor; *Proc. Natl. Acad. Sci. USA* **86** 3963–3967

Roeder R G 1991 The complexities of eukaryotic transcription initiation: regulation of pre-initiation complex assembly; *Trends Biochem. Sci.* **16** 402–408

Sawadogo M and Sentenac A C 1990 Polymerase B (II) and general transcription factors; *Annu. Rev. Biochem.* **59** 711–754

Schreiber E, Miller M M, Schaffner W and Matthias PC 1989 Octamer transcription factors mediate B-cell specific expression immunoglobulin heavy chain genes; in *Tissue specific gene expression* (*ed*.) R Renkawitz (Weinheim: VCH Verlagsgesellschaft) pp 33–51

Spencer C A, Lestrange R C, Novak V, Hayward W and Groudine M C 1990 The block to transcription elongation is promoter dependent in normal and Burkitts Lymphoma c-myc alleles; *Genes Dev.* **4** 75–88

Wasylyk B 1988 Enhancers and transcription factors in the control of gene expression; *Biochim. Biophp Acta* **951** 17– 035

Wirth T, Staudt L and Baltimore DC 1987 An octamer oligonucleotide upstream of a TATA motif is sufficient for lymphoid-specific promoter activity; *Nature* (*London)*) **329** 174–178