A MOLECULAR ANALYSIS OF THE FIVE PRIME END REGIONS OF THE MOUSE MAJOR URINARY PROTEIN GENES.

PETER GHAZAL

Thesis presented for the degree of Doctor of Philosophy

University of Edinburgh

1986

To my mother and father.

I declare that this work is my own, except where otherwise stated.

Peter Ghazal

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. John Bishop, for his help, advice, and patience throughout this work.

Many thanks to Ann Duncan, Melville Richardson and Morag Robertson for their technical assistance given towards the end of the project. Also, thanks to Iain McIntosh for typing the references and some of the tables; George Newell, and his staff at the workshop; Frank Johnson and Alan McEwan for the photographs.

In addition, I am grateful to many of my colleagues for advice and stimulating discussions. In particular, I thank Dr. John Clark, Dr. Sam Martin, Dr. Raya Al-Shawi, Dr. Paul Simons and Alistair Chave-Cox.

Last but not least, my thanks to Jacqueline for her moral support during the preparation of this work.

ABSTRACT

The mouse major urinary proteins (MUPs) are encoded by a multigene family of about 35 genes. The majority of the genes belong to one of two groups, Group 1 and Group 2. The predominant arrangement of these genes is a 45-kb repeat structure containing, in divergent linkage, a Group 1 and a Group 2 gene. MUP genes are differentially regulated by various hormones in a variety of tissues. Therefore, a comparison of the pattern of expression with the gene structure should allow their tissue-specific regulatory sequences to be identified.

A study of the 5' end regions of the MUP genes is presented. The start site of transcription of the most abundantly transcribed MUP genes (the Group 1 genes) was determined. On the basis of two criteria, nuclease S1 protection and primer extension analysis, the initiation site of transcription of the Group 1 genes is 31-bp downstream of the TATA box. To further characterize the 5' end regions of the genes the sequence structure of 8 MUP genes was collated from either the first intron or the second exon to about position -600. The implications of these data concern three important aspects of the structure and function of the genes. 1) Evolutionary implications: The sequenced Group 2 genes contain a common stop-codon within the first exon and have consequently arisen from a common ancestral pseudogene. The sequence data also suggest that homogenisation processes have occurred between members of a Group but not between Groups. Together these observations are in agreement with the proposal that the 45-kb unit is the basic unit of

MUP gene evolution. 2) Number of functional genes: The estimated number of Group 2 pseudogenes is about 12. Taking into account other MUP pseudogenes that have been described, the total number of functional MUP genes is estimated to be around 20. This number agrees reasonably well with the total number of different MUP proteins that are synthesised amongst the various tissues. 3) Identification of regulatory sequences: A simple short repetitive sequence mainly of A residues is situated upstream of position -50 in the MUP genes. This sequence varies strikingly in length and composition between the genes. This region contains the major sequence variation between the different MUP genes. Due to its proximity to the TATA box, it is proposed to have a major functional significance. A computer analysis of the 5' flanking regions of the MUP genes identified various symmetries and putative cis-acting regulatory sequence motifs. In particular, sites for the binding of steroid hormone receptors and heavy-metal regulatory sequence motifs were found.

Finally, an investigation of the promoter function of a Group 1 promoter in BHKtk- fibroblast cells was undertaken. Data is presented on the transient expression in BHKTk- cells of constructs containing various MUP promoter sequences linked to the HSV-Tk gene in the presence of either the SV40 enhancer, or the SV40 enhancer and early promoter sequences. These studies firmly demonstrate the enhancer activation of the MUP promoter in BHK cells. It is suggested that this enhancer dependence involves a deregulatory effect on the MUP promoter by the enhancer while the MUP promoter and associated sequences has a down regulatory effect on the SV40 early promoter region.

CONTENTS

Chapter 1: THE MUP GENE FAMILY

Introduction. 1

Sex Differences. 2

Temporal Expression. 2

Hormonal Regulation. 4

MUP Gene Organisation. 4

MUP Gene Structure. 6

Chromatin Structure. 7

Expression of MUPs in Hepatocytes. 8

Proteins Evolutionary Related to MUPs. 9

Aims of Project. 11

Chapter 2: INITIATION OF TRANSCRIPTION IN THE GROUP 1 GENES.

Introduction. 13

S1-Mapping and Primer Extension of Group 1 MUP mRNA from the BALB/c Mouse Male Liver. 14

An Alternative First Exon. 17

An Artefact from the S1 Protection Probe. 18

Conclusions. 21

Chapter 3: SEQUENCE STRUCTURE OF THE 5' ENDS AND EVOLUTION OF THE MUP GENES.

Introduction. 22

Sequencing Strategies. 26

Comparison of Group 2 Genes. 34

Transcriptional Signals. 34

Translational Signals. 35

Evolutionary Implications. 36

Comparison of Group 1 Genes. 39

Transcriptional Signals. 39

Translational Signals. 41

Evolutionary Implications. 41

Comparison of BS105A with a Group 1 and a Group 2 Gene. 45

Primary Structure. 45

Evoltionary Implications. 47

Conclusions. 49

Chapter 4: STRUCTURAL CHARACTERIZATION OF THE MUP PROMOTER REGION.

Introduction. 50

A Hypervariable Region in the MUP Promoter. 52

Possible Functions for the A-Tract Region

1. Opening (entry) site for RNA polymerase. 54

2. Positioning of regulatory elements. 54

3. A determinant for nucleosome phasing. 54

Repeats and Palindromes. 57

Group 1 Genes. 58

Direct Repeats. 58

Palindrome. 58

Group 2 Genes. 61

Direct repeats. 61

Inverted repeats. 62

Regulatory Sequence Motifs? 63

Glucocorticoid consensus sequences. 64

Enhancer core sequences. 67

Nuclear factor 1 binding sites. 68

Metal responsive elements. 71

Multiple Sequence Motifs. 73

Chapter 5: ENHANCER ACTIVATION OF THE MUP PROMOTER.

Introduction. 76

Expression of genes in tissue-culture cells. 76

Widely-expressed genes. 77

Inducible genes. 78

Tissue-specific genes. 78

Positive regulation. 80

Negative regulation. 80

The SV40 early promoter region. 81

Plasmid Constructions and Nomenclature. 84

Analysis of Constructs. 89

Tranlational Effects. 90

Promoter Utilization in BHKtk- Cells. 93

pSVEP. Constructs. 93

pSVE. Constructs. 95

Comparison of tk Activities.

Low efficiency of the MUP promoter. 99

Removal of the SV40 promoter. 99

Resection of the MUP promoter linked to the SV40 Enhancer. 100

Conclusions. 102

Chapter 6: METHODS

Enzymes. 103

Preparation of mouse male total mRNA. 103

Fractionation of poly(A)+ RNA. 104

Preparation of nuclease S1 and primer extension probes. 104

S1 mapping and primer extension of mRNA. 105

Cloning. 105

DNA sequencing. 105

Plasmid DNA preparations. 106

Tissue-Culture Methods. 106

Cells and media. 106

Transfection of BHKtk- cells 107

Thymidine kinase enzyme assays. 107

Nuclease S1 protection analysis of mRNA extracted from transfected cells. 108

REFERENCES 111

ABBREVIATIONS 138

APPENDIX 141

TABLES

1. Relative level of MUP mRNA in expressing tissues. 3

2. Comparison of the start site and TATA boxes of Group 1 and Group 2 genes with the consensus TATA box and cap site. 20

3. Cloning strategies 28

4. Frequency of base changes between the 5' sequences of the Group 1 genes. 42

5. A summary of regulatory sequence motifs found within the MUP promoter region. 65

6. A summary of tk assays during transient expression in BHK cells. 98

FIGURES

1. Nuclease S1 protection and primer extension analysis. 16

2. Artefacts from the S1 protection probe. 19

3. Schematic representation of sequencing strategies. 30

4. Schematic representation of sequencing strategies. 31

5. A representative example of a sequencing gel. 33

6. Sequence comparison of Group 2 genes. 37

7. Sequence comparison of Group 1 genes. 40

8. N-terminal sequences of BS6, MUP15 and BS105A. 44

9. Sequence comparison of BS105A with BS6 and BS2. 46

10. A-tract region. 53

11. Symmetries within the MUP promoter region. 60

12. Sequence structure of the MUP promoter region. 69

13. Schematic representation of the constructs. 85

14. Sequence structure of the 5' end region of the Group 1 gene BS6. 87

15. Sequence structure of the SV40 early promoter region. 88

16. Translation products of the 5' end of pSVEP.S2B(or S1B). Tk constructs. 91

17. S1 analysis of mRNA synthesised in BHK cells. 94

18. Northern analysis of mRNA synthesised in BHK cells. 96

19. Thymidine kinase activity from cell extracts prepared from transfected cells (Experiment 4). 109

THE MUP GENE FAMILY

INTRODUCTION

The major urinary proteins (MUPs) of the mouse are a family of closely related, small, acidic proteins that are synthesised in large quantities in the liver, secreted into the plasma and rapidly excreted into the urine. MUP genes are also expressed in the mammary (M), lachrymal (LM), submaxillary (SM), parotid (P) and sublingual (SL) glands, but at much lower levels than in the liver (Shaw et al, 1983). In vitro translation of liver mRNA shows that there are at least twelve different MUP species expressed in this tissue (Clissold and Bishop, 1982; Shaw et al, 1983; Shahan and Derman, 1984). The in vitro translation products of MUP mRNA from the SM, P, SL and M glands largely comprise different subsets of the liver products, while the LM gland mRNA gives rise to a different set of MUPs (Shaw et al, 1983). However, it is not known whether the LM products are members of the liver set which have been post-transcriptionally modified, or products of a subset of liver genes with a different mode of transcription, or products of genes that are not active in the liver. In total approximately 20 different MUPs from these various tissues can be distinguished. The MUPs are under multihormonal control and variation in hormonal responsiveness is detected between MUPs that are expressed in the same tissue as well as between MUPs that are expressed in different tissues. Thus the MUP genes are a highly tissue-specific family which are differentially regulated within and between different tissues.

Sex Differences

In BALB/c adult mice, MUP mRNA makes up about 8% of the total liver poly(A)+ mRNA, this level being five times higher than that of adult female mice (Hastie and Held, 1978; Hastie et al, 1979). Female mice show a simpler liver MUP pattern than male mice, although treatment with testosterone induces a male-like pattern (Finlayson et al, 1965; Szoka and Paigen, 1978; Clissold et al, 1984). Whether the testosterone has a direct effect on the liver MUP mRNA is not known, although work by Norstedt and Palmiter (1984) on the growth hormone regulation of MUP expression in the liver suggests that the gonadal steroids may exert their effects either on the hypothalamus or directly on the pituitary. The LM gland, like the liver, shows sexual dimorphism, with the male having approximately five times as much MUP mRNA as the female (Shaw et al, 1983). The SM gland does not show sexual dimorphism with respect to MUP expression.

Temporal Expression

The expression of MUP in different tissues is under different developmental control. In the liver, MUP mRNA is first detected in 3 week old male mice, full expression being reached only 6-7 weeks after birth. Derman (1981) showed that the different levels of MUP mRNA in the livers of mice of different ages and sex are reflected in differences in the rate of transcription. The LM gland has adult MUP mRNA levels at two weeks of age. This corresponds to the

TABLE .1. Relative level of MUP mRNA in expressing tissues

Tissue	Max Level of MUP mRNA (copies per cell)	Hormonal Regulation	Influencing Hormones	First Detectable Time of Expression
Liver (male)	30,000	+	T, T_4, GH	3 weeks
Lachrymal (male)	6,000	+	T	2 weeks
Submaxillary gland	1,250	-	none	1 week
Mammary gland	1,000	?	unknown	1st pregnancy

(From Shaw *et al.*, 1983).

earliest time at which lachrymal glands can be identified for dissection. The SM gland shows first detectable levels of MUP mRNA at one week of age, maximal levels being achieved between 4 and 7 weeks. MUP expression in the M gland is detected at the first pregnancy (Shaw *et al*, 1983, see Table 1).

Hormonal Regulation

The hormonal regulation of MUPs is different in the different tissues. Liver MUP mRNA is regulated by testosterone, thyroxine, growth hormone and, in some strains, glucocorticoids (Knopf *et al*, 1983; Norstedt and Palmiter, 1984). Using thyroidectomised, hypophysectomised female mice and mutant mice (*little* male and *tfm*/Y mice), Knopf *et al* (1983) have found that testosterone, growth hormone and thyroxine modulate MUP synthesis. Shaw *et al* (1983) have also found that different liver MUP components are regulated differently by testosterone, thyroxine and growth hormone. In the LM gland, testosterone induction of MUPs appears to be independent of growth hormone and thyroxine, while in the SM gland, MUP expression does not appear to be under hormonal regulation. The hormonal regulation of the mammary gland MUPs has not yet been studied. However, it is not known whether the hormonal modulation of MUPs is by direct or indirect action on the particular tissues (see section on the expression of MUPs in hepatocytes).

MUP Gene Organisation

The MUPs are encoded by a family of about 35 genes, tightly clustered on chromosome 4 (Bennett _et al_, 1982; Bishop _et al_, 1982; Krauter _et al_, 1982). Most of the genes fall into two main groups, Group 1 and Group 2. Between 13 and 15 Group 1 genes and the same number of Group 2 genes are arranged pairwise in head-to-head configuration about 15-kb apart, forming a set of gigantic imperfect palindromes each about 45-kb in size (Clark _et al_, 1984b). The symmetrical regions in each palindrome include the genes themselves and extensive parts of their 5' and 3' flanking regions (Bishop _et al_, 1985). The Group 2 member of each pair of genes is a pseudogene while the Group 1 member is functional (Ghazal _et al_, 1985). Thus the 13 to 15 45-kb palindromes contain 13 to 15 pseudogenes and the same number of active Group 1 genes. Between 5 and 9 genes do not fall into either Group 1 or Group 2 and appear not to be incorporated in palindromic structures. Of these at least 3 are pseudogenes (Clark _et al_, 1982; Clark _et al_, 1984b; Al-Shawi, 1985). Thus the total number of active MUP genes is about 20, roughly the same as the number of different proteins synthesised in the liver and lachrymal glands together.

Sequence homologies between the palindromes, on the basis of heteroduplex formation, hybridization and sequence data, suggest they have arisen from a common ancestor. This implies that the 45-kb palindrome is not only the predominant unit of MUP gene organisation but also the unit of MUP gene evolution. It has been proposed that the contemporary array of MUP genes has arisen through the spread of the 45-kb palidrome, by replacement of an ancestral array brought about by a process of unequal crossing-over (Ghazal _et al_, 1985).

Ten different Group 1 genes, four Group 2 genes and a MUP15 gene (a gene outwith both Group 1 and Group 2) have been isolated. A phylogenetic analysis based on restriction sites and small deletions or insertions in the genes and their flanking regions and on comparative sequence data indicates that the Group 1 genes fall into two main subgroups (Al-Shawi, 1985).

MUP Gene Structure

The transcription unit of MUP genes is 3.9-kb long and contains seven exons (Clark et al, 1984a). The first six exons contain the coding sequences, while the last exon consists entirely of non-coding sequences. Three different splicing configurations have been found, which result from the presence of alternative splice sites within the untranslated region of exon 6 (Clark et al, 1984a and Clark et al, 1985b). The most abundant liver MUP transcripts contain part of exon 6 and all of exon 7. The less abundant (by a factor of about 10) and smaller liver MUP transcripts contain an extended exon 6 but completely lack exon 7. On the basis of hybridization most of the long transcripts belong to the Group 1 genes while the shorter transcripts are predominantly derived from the Group 2 genes (Clark et al, 1984a).

The mRNA-specifying sequence of four Group 1 genes and almost full length sequences of four Group 1 cDNAs have been determined. These code for a signal peptide 18 amino acids long, and a mature protein 162 amino acids long. Although nucleotide homology between the different Group 1 sequences is on average

99.7%; they nevertheless specify different proteins. At present it is not known which genes code for which protein. MUPs generally have been found not to be glycosylated (Szoka and Paigen, 1978), although the sequence of MUP15 (cDNA) specifies a significantly different (31%) amino acid sequence, which in particular contains a potential N-linked glycosylation site. There is direct evidence that MUP15 specifies a minor glycosylated MUP protein in the urine of mice (Clark et al, 1985b; Kuhn et al, 1984).

The transcription units both of a Group 1 gene (BS6) and of a Group 2 pseudogene (BS2) have been sequenced. The sequence homology between the two genes is about 90%. The 5' flanking regions contain distinctive TATA boxes, and in general ~~in general~~ are very similar except that about position -50 to the cap site the Group 1 gene contains a long A-rich sequence which is replaced in the Group 2 pseudogene with a much shorter sequence. The upstream regions of five Group 1 and four Group 2 genes have also been sequenced (this Thesis). The major differences observed are in the length and composition of this A-rich tract. The upstream regions contain sequences that show homologies to sites known to bind trans-acting regulatory factors (this Thesis).

Chromatin Structure

The location of DNase I hypersensitive domains within the 45-kb duplication unit have been mapped in the liver and kidneys of male and female mice at various stages of development (J. Clark, unpublished results). No sites are developed in the kidneys, a tissue that does not express MUP. Ten hypersensitive sites are fully established in the liver of three week old mice, the time at which MUP mRNA synthesis is first observed. No differences are seen between male and females. Eight of the hypersensitive sites are arranged in similar positions around the Group 1 and Group 2 genes. These are present 0.5-kb 3' to the polyadenylation site and 0.75, 2.25 and 7 kb 5' to the cap site of the genes. Two non-symmetrical hypersensitive sites are present 0.5 and 5.5 kb 5' to the cap site of the Group 2 genes. Nuclease hypersensitve domains, a characteristic of active chromatin, are believed to be regions of nucleosome-free DNA (Elgin, 1984) which are involved in interaction with specific trans-acting factors (Emerson et al, 1985). For this reason it is thought possible that these sites are involved in the liver-specific regulation of MUP gene expression.

Expression of MUPs in Hepatocytes

When liver cells are dissociated and plated (hepatocytes) transcription of the MUP genes is rapidly switched off. This results in a dramatic fall in the level of MUP mRNA to virtually zero in about four days (T. Spiegelberg, unpublished results). During the same time period the level of transferrin mRNA rises by about two-fold and alpha-fetoprotein mRNA appears de novo, showing that the integrity of the cells is maintained in culture during changes in expression which seem to mimic liver regeneration. It has been shown that growth hormone, thyroxine and insulin all retard the rate of decay of MUP transcription in cultured hepatocytes, and that growth hormone and thyroxine are synergistic (T. Spiegelberg, unpublished results). This suggests that growth hormone, thyroxine and insulin directly effect the expression of MUPs in the liver.

Proteins Evolutionarily Related to MUPs

In the rat, a homologous gene family codes for the alpha-2u globulins (Kurtz, 1981a; Dolan et al, 1982). Alpha-2u globulins are synthesised in the liver of male rats (Laperche et al, 1983). Hepatic alpha-2u globulins are regulated by thyroxine, testosterone, glucocorticoids, growth hormone, insulin and oestrogen, unlike submaxillary alpha-2u globulin which does not appear to be under hormonal regulation (Motwani et al, 1980; Lynch et al, 1982; Roy et al, 1983; Laperche et al, 1983; Kulkarni et al, 1985). Dexamethasone-induced expression of these genes following their introduction into mouse fibroblast cells has been reported (Kurtz, 1981b). However, there are reasons to believe that difficulties have been encountered in confirming this data.

The alpha-2u globulins are encoded by a multigene family of about twenty genes (Kurtz, 1981b). Comparison of a rat alpha-2u globulin gene and a mouse Group 1 gene showed that the transcription units are similar in structure and that the exonic sequences are about 80% homologous. Furthermore, replacement sites (those at which mutations alter the amino acid that is coded) have occured less frequently between the genes than silent sites (those at which mutations do not change the amino acid that is coded). This suggests that the protein sequences are being conserved, while the genes are rapidly evolving (Clark et al, 1984a).

Recently it has been found that MUP genes belong to a gene superfamily. Significant homologies have been detected between MUP and alpha-2u globulin; beta-lactoglobulin, a secretory protein found in the milk of ruminants; alpha-1 microglobulin, a low molecular weight human serum protein; human retinol-binding protein, a serum protein that binds retinoic acid; and human alpha-1-acid glycoprotein, a serum protein (Pervaiz and Brew, 1985; J.O. Bishop, unpublished).

As yet the function of the mouse MUPs is not known, although there is strong circumstantial evidence which suggests they may be involved in behavioural communication (Vandenbergh et al, 1976 and Shaw et al, 1983). Furthermore it has been proposed that the active component of these pheromonal effects mediated by MUPs is the first six N-terminal amino acids of the proteins (Clark et al, 1985b; see Appendix). Experiments are currently in progress to ascertain whether or not this model is correct. However, preliminary observations in this laboratory indicate that MUPs may be induced by exposure of heavy-metals (J. Whitaker, unpublished and this Thesis). The corollary to this is that MUPs may be involved in heavy-metal detoxification. These observations together raise the possibilty that the MUPs are bifunctional.

AIMS OF PROJECT

The project described in this thesis is concerned with the characterization of the 5' end region of the MUP gene family. The 5' end regions of most if not all eukaryotic structural genes that have been studied are known to be necessary for expression. As discussed previously different members of the MUP gene family show temporally and hormonally different modes of expression within specific tissues. The main objective of this study is to explore the structure of the 5' flanking region of the MUP genes in relation to their regulated expression.

As a first step, the start point of transcription of the most abundantly transcribed MUP genes (the Group 1 genes) was determined. Once this region had been defined the 5' flanking sequences of a number of MUP genes were collated. In the long term it is hoped that the protein products of various cloned MUP genes will be identified, or at least that their transcriptional products will be identified so that a direct comparison between their hormonal modulation and/or tissue-specific expression may be made with their 5' end sequence structure. Since the members of the MUP gene family are structurally very homologous any differences seen at their 5' end will be significant. The corollary of these data has also lead to important observations regarding the structure and evolution of the MUP gene family. For example, it was found that most if not all the Group 2 genes are pseudogenes.

Since it has not been possible to distinguish the various MUP transcripts or to assign the cloned genes to their protein products, two approaches have been used to investigate the putative regulatory nature of this region. Firstly, a computer search of the 5' end sequence data for symmetries and known _cis_-acting regulatory motifs was done. The second approach was the analysis of a MUP Group 1 gene promoter region in tissue-culture cells. This involved the transient expression of various constructs containing the MUP promoter region linked to the SV40 early control region. This approach demonstrated that the MUP Group 1 promoter is functional under the control of the SV40 enhancer. Furthermore, these expression studies demonstrated the utility of the constructs for the establishment of transgenic mice for the analysis of MUP gene expression.

INITIATION OF TRANSCRIPTION IN THE GROUP 1 GENES.

INTRODUCTION

The synthesis of messenger RNA in eukaryotes is carried out by RNA polymerase II. The initiation of transcription occurs at a discrete site(s) on the DNA, the start site or cap site. According to the cap-promoter hypothesis of Ziff and Evans (1978), the 5' end of the mRNA delineates the start of transcription and therefore the proximity of the promoter. Evidence for this comes from the fingerprint analysis of the 5' end of nascent transcripts and is also supported by the S1 mapping of precursor and mature mRNA (Weaver and Weissmann, 1979; Bunick et al, 1982). These 5' termini are modified with a 'cap' structure m7G(5')ppp(5')N. Recently Konarska et al (1984) have suggested that the cap structure plays an important role in splicing, since the addition of cap analogs to HeLa whole-cell extracts inhibits *in vitro* splicing of mRNA. The exact role of the cap structure in splicing is not known, although it does not appear to be related to mRNA stability.

At the start point, there is no extensive homology of sequence, but there is a tendancy for the first base of mRNA to be an A, flanked on either side by pyrimidines (Breathnach and Chambon, 1981, Table 2). This consensus was collated from 22 genes and therefore due to the low number of genes compared is most probably biased. An almost ubiquitous sequence of 7-bp (TATAWAW), the TATA box, is located between 19 and 34-bp upstream of the start site. (Table 2).

There are only a few exceptions known, for example, the HMG CoA reductase gene (Reynolds *et al*, 1984) and the SV40 Late Promoter (Brady *et al*, 1982). In most of these exceptional genes multiple initiation sites are present. Deletion of the TATA box usually results in transcriptional initiation at many points besides the cap site. This has lead to the proposal that the TATA box functions as a selector for the initiation site of transcription. Deletion of the TATA box or point mutations within it have also been shown to effect transcriptional efficiency (Dierks *et al*, 1983 and Grosveld *et al*, 1982). The SV40 early region promoter seems to be an exception, for although deletion of its TATA box results in transcriptional initiation from multiple sites, transcriptional efficiency is not affected (Benoist and Chambon, 1981). This probably relates to the strong enhancing effect of the 72-bp repeats.

In general the transcriptional organisation of polymerase II genes follows the 'one promoter one gene' rule. However, in some genes initiation of transcription may occur from alternate promoters, for example the discoidin-I gene of *Dictyostelium* (Jellinghaus *et al*, 1982); the alpha-amylase-I gene of the mouse (Young *et al*, 1981); and the human myosin gene (Nabeshima *et al*, 1984).

Here, I have determined the precise point of initiation of transcription for the Group 1 MUP genes by mapping the 5' termini of Group 1 transcripts from the male mouse liver.

S1-MAPPING AND PRIMER EXTENSION OF GROUP 1 MUP mRNA FROM THE BALB/c MOUSE MALE LIVER.

On the basis of two criteria, S1 nuclease protection and primer extension, the Group 1 mRNA cap site is located 31 ±1-bp downstream from the TATA box (Figure 1). The S1 protection probe was a 696-bp *Alu*I fragment, extending from nucleotide +127 to nucleotide -568 in the MUP clone BS6 (Group1 gene) and cloned into the *Hinc*II site of M13mp7. The single-stranded M13 clone was annealed to the sequencing primer and the strand complementary to MUP mRNA was uniformly labelled using the Klenow fragment of DNA polymerase I. The probe was excised and gel purified from the vector by digestion with *Eco*RI, retaining 15-bp of vector polylinker sequences on each side of the probe. This was hybridized to Poly(A)+ RNA from BALB/c mouse male liver and challenged with S1 nuclease. Three protected fragment lengths of 127, 130 and 142 nucleotides were observed (Figure 1). The primer extension probe was the 93-bp *Alu*I - *Sau*96I fragment between nucleotides +127 and +34 in the BS6 sequence. This was prepared and annealed to poly(A)+ RNA in essentially the same way as the S1 protection probe and complementary DNA was synthesised using AMV reverse transcriptase. One major extended product of 127 nucleotides is seen with the liver mRNA (Figure 1). Interestingly, the next largest band to 127 is a minor product of the reaction (122 nucleotides) which is also present in all of the control tracks and is most likely an artefact generated from the self extension of the primer. Such anomalies of the primer extension method have been reported before (Lee and Roeder, 1981).

FIGURE 1.

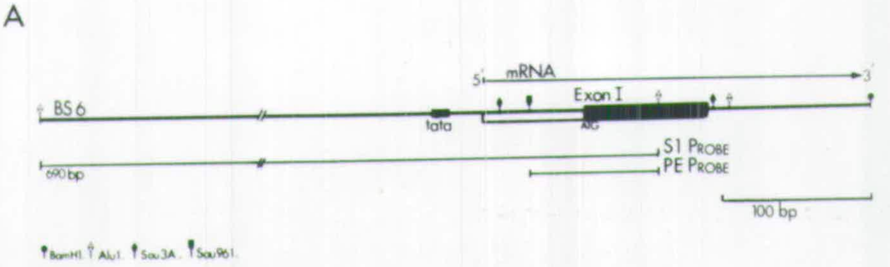
S1 protection and primer extension.

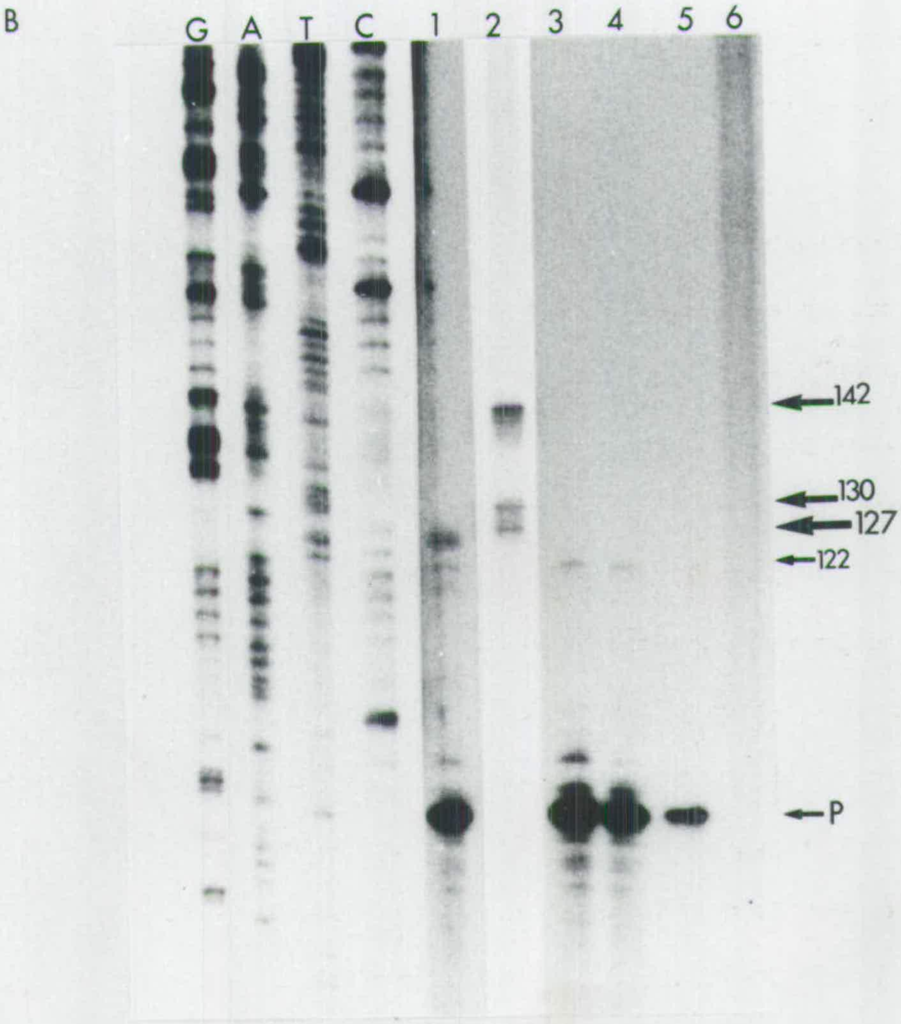
A, Restriction map of the 5' region of MUP BS6 showing its relationship to the probes used for S1 protection and primer extension. Open and closed boxes show the untranslated regions and translated regions of exon 1.

B, Electrophoretic analysis of the products of S1 protection and primer extension. Lanes G-C , sequence ladder of the S1 probe used to provide MW markers; Lane 1, primer extension of liver poly(A)+ RNA; Lanes 3-5, primer extension controls- kidney poly(A)+ RNA (3), no RNA (4), primer extension probe alone (5); Lane 6, S1 protection of kidney poly(A)+ RNA. P is the primer extension probe.

In the two tracks with liver poly(A)+ RNA both the S1 analysis and the primer extension analysis yield bands 127-128-bp long which positions the mRNA cap site 31 ± 1-bp downstream from the TATA box. In lane 2 bands 142 and 130-bp are observed and are discussed in the text.

C, Sequence of part of exon 1 and flanking sequences showing the TATA box, cap site and the various restriction cuts used for probe preparation.





C

```
      -31                            +1 Leader                     Sau96I
gagagtatataaggacaagcaaaggggctggggagtGGAGTGTAGCCACGATCACAAGAAAGACGTGGTCCT

                              Met
GACAGACAGACAATCCTATTCCCTACCAAAATGAAGATGCTGCTGCTGCTGTGTTTGGGACTGACCCTAGTC

                 AluI
TGTGTCCATGCAGAAGAAGCTAGTTCTACGGGAAGCAA
```

The primer extension probe and S1 protection probe have the same AluI cut at position +127 of BS6. Therefore, the extended and protected products should have the same fragment length. The only common fragment is the 127 nucleotide band which therefore probably delineates the initiation site of transcription. The 142 and 130 nucleotide bands are exclusive to the S1 mapping experiment. Careful consideration of the differences between the two techniques suggested two ways in which these bands could have arisen. The first is the possibility that they represent an alternative upstream exon. Alternatively, they may be artefacts generated from the S1 protection probe.

An Alternative First Exon

Since the S1 protection probe is uniformly labelled it is possible for it to detect an alternative first exon within its fragment length. This explanation would imply that the 127 nucleotide band represents a downstream exon 1 the sequence of which is complementary to both the primer extension probe and the S1 protection probe. The 142 and/or 130 nucleotide bands would represent an upstream exon 1 the sequence of which is complementary only to the upstream region of the S1 protection probe. This possibility could be tested either by using an end-labelled S1 protection probe, in which case only the 127 nucleotide band should be developed, or by using a uniformly-labelled probe that lacks the region of the downstream exon. In this case only the 142 and/or 130 nucleotide bands should be developed. The latter experiment was performed using a probe extending from the Sau96I site at +34 to the AluI site at -568.

The result of this experiment was negative (data not shown). However, it must be noted that the interpretation of a negative result is only that it does not offer any positive evidence for the presence of an alternative exon.

An Artefact From The S1 Protection Probe

The second and less attractive explanation is that the 142 and 130 nucleotide bands are artefacts produced by sequence homology between the 15 nucleotide polylinker extension at the 5' end of the probe and the MUP mRNA. Some homology is in fact present and is shown in Figure 2.C. Although the homology is present, it may not explain the observed results. Whether or not it does so can easily be tested by removing the polylinker sequences from the S1 protection probe. This was done by cutting the probe with BamHI leaving only 8 nucleotides of polylinker sequence attached to the 5' end of the probe. When the S1 protection experiment was repeated using the BamHI-cut probe, the 127 and 130 nucleotide bands were found but the 142 nucleotide band was absent (Figure 2.A). Thus the 142 nucleotide band seems to be artefactual.

The 142 and 130 nucleotide bands are more susceptible to S1 nuclease attack than the 127 nucleotide band as shown in Figure 2.B. This suggests that not only the 142 but also the 130 nucleotide band is an artefact generated from the S1 protection probe, and that there is a single transcription initiation point corresponding to the 127 nucleotide band.

FIGURE 2.

Artefacts from the S1 protection probe

A and B, Electrophoretic analysis of S1 protection probe. Lanes 1-4 and 1'-4', sequence ladders of previously characterized M13 clones used as MW markers.

A: Lane RI, S1 protection of liver poly(A)+ RNA with EcoRI cut S1 probe; Lane HI, S1 protection with BamHI cut S1 probe; Lane CI, control for S1 protections - no RNA.

B: Lanes A-E, S1 protection of liver poly(A)+ RNA with EcoRI cut S1 probe, challenged with increasing amounts of S1 nuclease. 100 u/ml (A), 250 u/ml (B), 500 u/ml (C), 750 u/ml (D) and 1000 u/ml (E).

These show that the 142 and 130 bp bands are artefacts which result from partial homology of the MUP mRNA sequences immediately 3' to the AluI sequence to the polylinker region of M13 that is present in the S1 probe.

C, The probable sequence alignment between the MUP mRNA and the polylinker sequences of the S1 probe cut with EcoRI or BamHI. The polylinker overhang is underlined and the arrows indicate the positions of the S1 protection products.

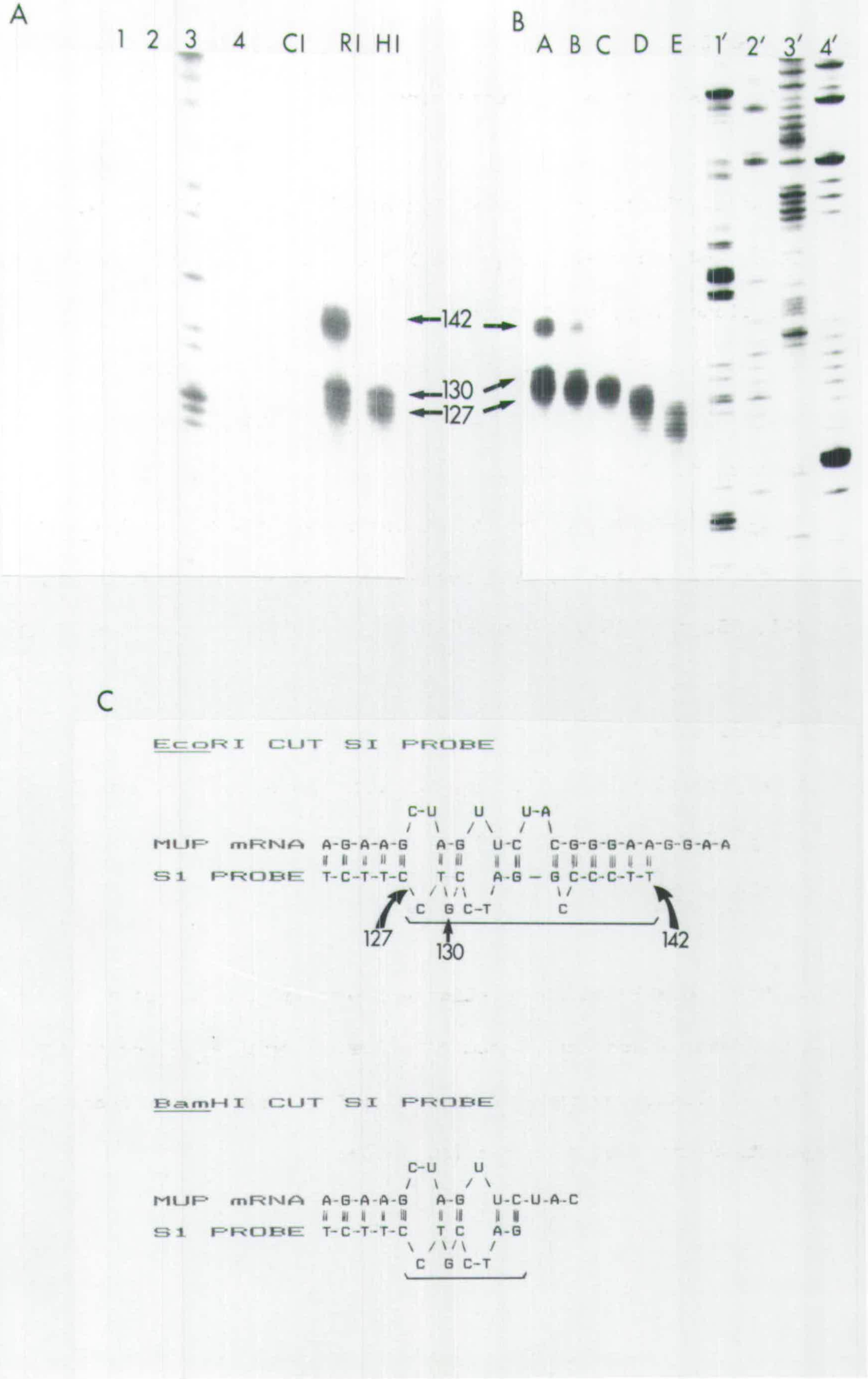

A
1 2 3 4 CI RI HI
B
A B C D E 1′ 2′ 3′ 4′
142
130
127
C
EcoRI CUT SI PROBE
C-U U U-A
MUP mRNA A-G-A-A-G A-G U-C C-G-G-G-A-A-G-G-A-A
S1 PROBE T-C-T-T-C T C A-G — G C-C-C-T-T
C G C-T C
127
130
142
BamHI CUT SI PROBE
C-U U
MUP mRNA A-G-A-A-G A-G U-C-U-A-C
S1 PROBE T-C-T-T-C T C A-G
C G C-T

Table 2 . Comparison of the start site and TATA boxes of group 1 and group 2 MUP genes with the consensus TATA box and cap site.

Position		1		3	4	5	6	7	8	9	10		12			15							+1					
Consensus from 60 genes		G	-	G	T	A	T	A	W	A	W	-	G	-	-	G	↔	Py	-	-	-	Py	A	Py	Py	Py	Py	Py
(Breathnach & Chambon, 1981)																	9-17bp											
	A	10		8	4	58	4	51	38	53	30	20	11			14		2				3	21	1	6	3	3	4
Base	T	10		6	49	1	56	6	22	6	20	7	9			10		11				7	0	7	7	6	8	11
Frequency	G	30		32	1	1	0	0	0	0	8	23	29			26		3				1	1	3	1	0	3	3
	C	9		14	6	0	0	3	0	1	2	10	11			10		6				11	0	11	8	13	8	4
BS6, BS5, BL1, BS1,		G	A	G	T	A	T	A	T	A	A	G	G	A	C	A	↔	G	-	-	-	T	G	G	A	G	T	G
BS102-2, BL25																	14bp					Group 1 cap site						
BS109-1		G	A	T	T	A	T	A	T	A	A	G	G	A	C	A												
BS2		G	A	G	T	A	T	A	T	G	A	G	G	A	C	A												
BS109-2		T	A	G	T	A	T	A	T	A	A	G	G	A	C	A												

CONCLUSIONS

The 127 nucleotiode fragment, common to both the S1 and primer extension analysis, maps the 5' end of the Group 1 genes to 100-bp upstream from the initiation codon and 31-bp downstream from the TATA box (Figure 1.C). Nine sequenced Group 1 genes (this Thesis and R. Al Shawi, 1985) are nearly all identical in this region. Hence, this transcriptional start site is likely to be representative of all the Group 1 genes. It is also positioned in the same region as the cap site of the rat genes homologous to the MUPs, the alpha-2u globulins. The sequence around the Group 1 cap site bears little resemblance to the consensus cap site of Breathnach and Chambon (1981), see Table 2.

In addition this study has demonstrated the importance of using two independant techniques in determining the point of initiation of transcription.

SEQUENCE STRUCTURE OF THE 5' ENDS AND EVOLUTION OF THE MUP GENES.

INTRODUCTION

Many eukaryotic genes are members of gene families. Gene families constitute evolutionarily related genes that share common, although not necessarily identical, functions. Gene families vary considerably in size. They may be composed of two to three genes or several hundred genes. The size of a gene family may also vary considerably between different species. For example, the histone gene family of Xenopus contains 20 to 50 copies of its histone genes while the newt contains 600 to 800 copies (Hentschel and Birnstiel, 1981). The members of a gene family may be linked and/or dispersed on different chromosomes.

Some members of some gene families are pseudogenes. These are presumed to be evolutionary relics of once intact genes, which become selectively neutral with the occurence of a first deleterious mutation, for example, the rabbit pseudo-beta 2 globin gene (Lacy and Maniatis, 1980) and the human pseudo-alpha 1 globin gene (Proudfoot and Maniatis, 1980). Another type of pseudogene is the so-called processed pseudogene. These lack the introns and promoter region found within other members of the family (Lee *et al*, 1983; Karin and Richards, 1982; Dudov and Perry, 1984). It has been suggested that such pseudogenes arose from retroviral reverse-transcription of mRNA into cDNA and its insertion into the germ line.

Members of a gene family often show homology in their flanking sequences. Barring deletions and disruptions caused by insertions, the extent of these homologies can be presumed to depend on the length of the segment of DNA initially duplicated. Linked genes produced by duplications or inversions are sometimes further amplified together. For example, this sequence of events seems to underly the organisation of the silkmoth chorion gene family (Jones and Kafatos, 1980; Iatrou and Tsitilou, 1983) and the amplification of the MUP Group 1 and Group 2 genes (Clark et al, 1984b; Bishop et al, 1985).

The degree of homology between members of a gene family can be remarkably high within a species, while at the same time the same gene family shows substantial variation between species. For this reason it is thought that mechanisms must exist which lead to the homogenization of gene families within species. Two mechanisms of this sort have been proposed. One of these is unequal crossing-over (Smith, 1976). Multiple occurrences of unequal crossing-over would lead to sequence homogeneity because stochastically one nucleotide sequence would come to dominate the family. The seperate series of amplification events that occur in different related species can result in the replacement of an ancestral array by different sets of genes in each species. Another mechanism which may lead to gene homogenization is gene conversion. Gene conversion is non-reciprocal exchange between two homologous sequences. The phenomenon occurs when DNA strands from two allelic, or non-allelic but homologous genes, form a heteroduplex and correction of mismatched bases takes place. The occurrence of gene conversion was first recognised in fungi at the genetic level between alleles on homologous chromosomes (Radding, 1978). Subsequently non-allelic conversion events were shown to occur between members of a gene family on sister chromatids, on the same chromatid or on different chromosomes (Jackson and Fink, 1981; Klien and Petes, 1981; Scherer and Davis, 1980). These non-allelic conversion events are more significant in giving rise to the maintenance of sequence homogeneity among members of a family rather than allelic conversions. In mammals it has been suggested that non- allelic gene conversion has occured between globin genes (Slightom et al, 1980), mouse MHC genes (Weiss et al, 1983; McIntyre and Seidman, 1984) and immunoglobulin genes (Bentley and Rabbitts, 1983; Ollo and Rougeon, 1983).

Limited sequence data often does not allow one to distinguish between the products of gene conversion and unequal crossing-over. Consequently, the term 'gene conversion' has often been loosely used to describe genetic exchange between non-allelic genes, where change in sequence length is not observed and where all the products of a single recombination event are not known. For example, sequence comparison of the H-2K^{b} gene with a mutant allele H-2K^{bml} showed a total of seven nucleotide changes which are clustered in a 13 nucleotide region of the H-2K^{bml} gene (Weiss et al, 1983). Weiss et al (1983) suggested that these mutations have been introduced into the H-2K^{b} gene by gene conversion-like events. However, since there is limited sequence data available for additional mutant and other H-2 genes they cannot exclude the possibility of an unequal crossing-over event with a non-allelic gene. In addition, it is not formally excluded that the seven nucleotide changes seen are independent point mutations, or a hot spot of nucleotide substitutions. Nevertheless, since a H-2 class I (H-2L^d) gene contains the amino acids found in the bml sequence in this region, it is considered that a gene conversion event is the most likely cause for this phenomenon, in which case the H-2L^d is the likely donor gene.

In this chapter I describe the sequence structure of the 5' region of various members of the MUP multigene family, and relate the data to the evolution of the family. Part of this chapter is presented in the Appendix (see papers Ghazal et al, 1985 and Clark et al, 1985a).

SEQUENCING STRATEGIES.

DNA sequencing has become one of the most important tools for the analysis of genes. There are two major methods for DNA sequencing, the Maxam and Gilbert method and the chain terminator method of Sanger. The former method is preferably used for small scale sequencing and is based on the base-specific cleavage of the DNA by chemical reagents. The chain terminator method is best used for large scale sequencing and capitalises on two properties of the Klenow fragment of DNA Polymerase I. First, its ability to synthesize faithfully a complementary copy of a single-stranded template. Second, its ability to use 2' 3' dideoxynuleoside triphosphates as substrates. Once the 'dideoxy'analogue is incorporated the 3' end lacks a hydroxyl group and no longer is a suitable substrate for chain elongation; thus the growing DNA chain is terminated. Single-stranded DNA templates can easily be prepared from the single-stranded (ss) bacteriophage M13.

In this study the sequencing of the 5' end region of eight different MUP genes was performed by the chain terminator method. A total of about 11-Kb of sequence data was thus collated. A brief description of the various strategies used will be given and why the chain terminator method was chosen will be discussed.

Specific restriction fragments from lambda bacteriophage genomic clones or plasmid subclones of MUP genes (BS105A, BS109A, BS102A, BL25, BS1, BL1, and BS5) were cloned into unique restriction sites of the replicative form (RF) of M13 mp8, mp9 and tg131, genetically engineered versions of the ss bacteriophage M13. A summmary of the cloning strategies is presented in Table (3) and a schematic representation of the region sequenced is shown in Figures 3 and 4.

The sequencing of the 5' end regions of the genes was performed primarily on one strand. A maximum of 450 nucleotides may be read from one sequencing reaction with a combination of gradient and 'long run' gels (Biggin et al, 1983). Therefore, employing the strategy of 'clone-turn around' that is, cloning the restriction fragment in both orientations, the sequencing of a fragment as large as 800-bp is greatly facilitated (Winter and Fields, 1980). This is by far the simplest and quickest strategy for sequencing specific regions and was ideally suited for this study. A representative example of a sequencing gel is shown in Figure 5.

Where an insert is larger than 800-bp but less than 1.5-kb the remaining unsequenced region may be obtained by restriction with an appropriate 4-bp cutter and randomly selected DNA fragments cloned into M13 (the 'shot-gun method'). The limitation of this approach is the presence of appropriately sized restriction cuts and the number of fragments generated. The first limitation relates to the non-randomness of DNA sequences and the second to the selective screening potential for the desired clone. If a cloned fragment is already available from one gene then the homologous clone from other members of the family may be easily identified by hybridization to the existing clone and observing a shift in mobility of the hybrid on an agarose gel (Winter et al, 1981). However, selective

TABLE 3.

CLONING STRATEGIES

GENE	SUBCLONE	READING OF SEQUENCE STRAND	RESTRICTION CUT ENDS OF FRAGMENT	SIZE (bp)	M13 VECTOR	RESTRICTION CUTS OF VECTOR
BS109-2						
	109RPa	Sense	*EcoRI/PvuII*	1125	mp9	*SmaI/EcoRI*
	109PRb	Sense	*EcoRI/PvuII*	535	mp8	*SmaI/EcoRI*
	109RPc	Antisense	*EcoRI/PvuII*	535	mp9	*SmaI/EcoRI*
	109PRd	Antisense	*EcoRI/PvuII*	1125	mp8	*SmaI/EcoRI*
	109Sa	Sense	*Sau3A*	431	mp8	*BamHI*
	109Sb	Sense	*Sau3A*	140	mp8	*BamHI*
	109Sc	Antisense	*Sau3A*	140	mp8	*BamHI*
	109Sd	Antisense	*Sau3A*	335	mp8	*BamHI*
	109Se	Antisense	*Sau3A*	440	mp8	*BamHI*
BS102-2						
	102RPa	Sense	*EcoRI/PvuII*	1123	mp9	*SmaI/EcoRI*
	102PRb	Sense	*EcoRI/PvuII*	541	mp8	*SmaI/EcoRI*
	102RPc	Antisense	*EcoRI/PvuII*	541	mp9	*SmaI/EcoRI*
	102PRd	Antisense	*EcoRI/PvuII*	1123	mp8	*SmaI/EcoRI*
	102Sd'	Sense	*Sau3A*	335	mp8	*BamHI*
BL25						
	BEO	Antisense	*EcoRI/BamHI*	1279	mp8	*EcoRI/BamHI*
	E25	Antisense	Hong	1114		
	E17	Antisense	Hong	943		
	E15	Antisense	Hong	439		
	E14	Antisense	Hong	375		
	E12	Antisense	Hong	255		
BS105A						
	105RBa	Sense	*BamHI/EcoRI*	535	mp9	*BamHI/EcoRI*
	105BRb	Sense	*BamHI/EcoRI*	1125	mp8	*BamHI/EcoRI*
	105RBc	Antisense	*BamHI/EcoRI*	1125	mp9	*BamHI/EcoRI*
	105BRd	Antisense	*BamHI/EcoRI*	535	mp8	*BamHI/EcoRI*
	105Sa	Sense	*Sau3A*	160	mp8	*BamHI*
	105Sc	Sense	*Sau3A*	80	mp8	*BamHI*
BS1						
	H1	Sense	*HaeIII*	153	mp9	*SmaI*
	H2	Sense	*HaeIII/BamHI*	1140	mp9	*SmaI/BamHI*
	X	Sense	*XmnI/BamHI*	750	tg131	*EcoRV/BamHI*
	E1H	Antisense	*BamHI/HindIII*	2400	mp9	*BamHI/HindIII*
BL1						
	HF1	Sense	*HinfI**	490	mp8	*SmaI*
	E1H	Antisense	*BamHI/HindIII*	2400	mp9	*BamHI/HindIII*
BS5						
	H3	Sense	*HaeIII/BamHI*	775	mp9	*SmaI/BamHI*
	E1H	Antisense	*BamHI/HindIII*	2400	mp9	*BamHI/HindIII*
BS109-1						
	109-1RH	Antisense	*EcoRI/HindIII*	2200	mp9	*EcoRI/HindIII*

* blunt ended

screening procedures depend on the availability of subclones and therefore, without specific probes screening can be tedious. BS105A, BS102-2, BS109-2, BS109-1, BS1, BL1 and BS5 were sequenced by using one or a combination of the above strategies.

Where an insert is larger than 1-kb and no suitable restriction sites are available then a systematic sequencing strategy is best applied. (Poncz et al, 1982; Hong, 1982; Barnes et al, 1983; Henikoff, 1985). Such strategies are ideally suited for the chain terminator method where progressively shortened fragments are generated from the cloned insert. BL25 was sequenced in such a manner using the method devised by Hong (1982). This involves randomly nicking the RF with DNaseI in the presence of Mn++ to produce linear molecules. These are excised from an agarose gel, cut with a suitable restriction enzyme, religated and transfected. The sequence of each of these clones will originate at a random site within the insert. The second generation clones are screened by T tracking (that is only the chain termination for the T residues are analysed) and the choice of progressively shortened clones sequenced. This procedure requires a number of enzymatic steps, gel purification and a polyethylene glycol precipitation stage. Experience in this laboratory and others (Henikoff, 1985) have found variable success with this method. Particular attention must be given to the polyethylene glycol precipitation stage where variable recoveries are the most likely cause of failure. More recently a method based on the unidirectional digestion of the insert with Exo III has been devised by Henikoff (1985). This procedure requires no gel fractionation or purification step and therefore avoids many of

FIGURE 3 AND 4.

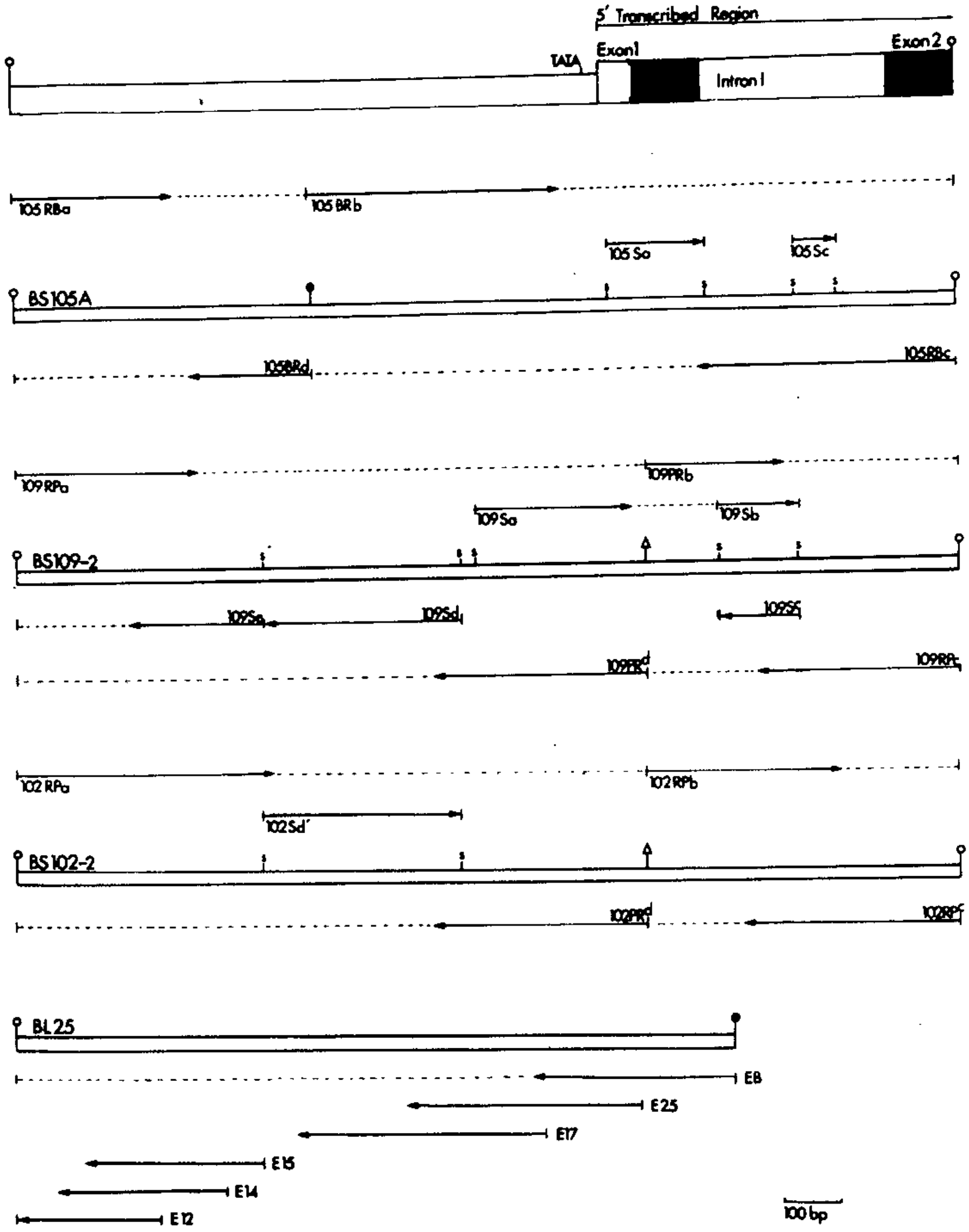
Schematic Representation of Sequencing Strategies.

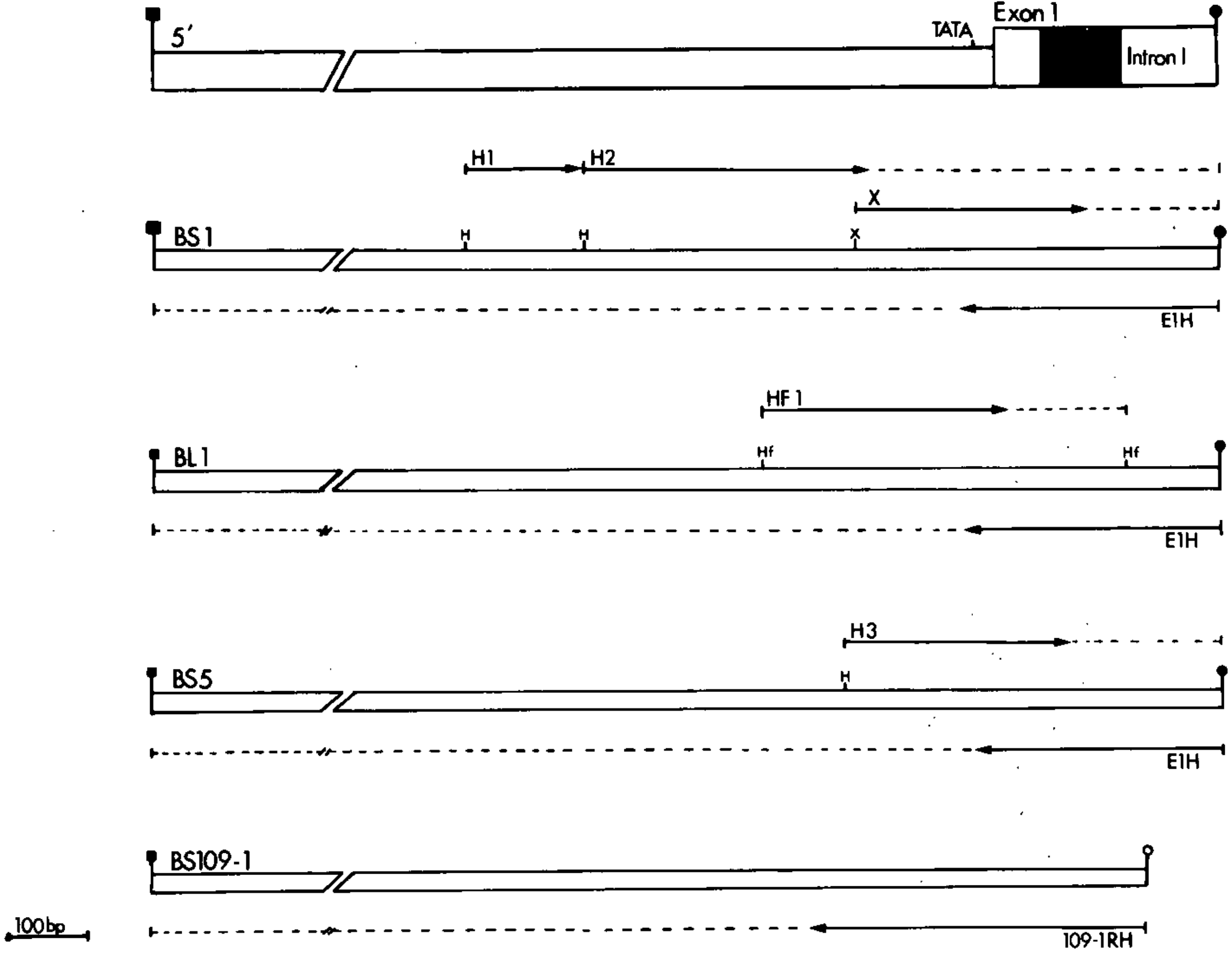
Fig. (3) shows the stategies applied to BS105A and the Group 2 genes while Fig. (4) shows the stategies applied to the Group 1 genes.

Arrowed lines indicate the sequenced regions and the dashed lines the unsequenced regions of the subclones.

Black boxes refer to the coding sequences of the genes.

The restriction map of each of the genes covers the region sequenced and shows the sites employed for the M13 cloning; ⚲, *Eco*RI; ↑, *Pvu*II; ■, *Hind*III; ●, *Bam*HI; s, *Sau*3A; H, *Hae*III; Hf, *Hinf*I; x, *Xmn*I.


5' Transcribed Region
TATA
Exon1
Exon 2
Intron I
105 RBa
105 BRb
105 Sa
105 Sc
BS105A
105BRd
105RBc
109 RPa
109PRb
109 Sa
109 Sb
BS109-2
109Se
109Sd
109Sf
109PRd
109RPc
102 RPa
102 RPb
102Sd'
BS102-2
102PRd
102RPf
BL25
E8
E25
E17
E15
E14
E12
100 bp


5′
TATA
Exon 1
Intron I
H1
H2
X
BS1
H
H
X
E1H
HF1
BL1
Hf
Hf
E1H
H3
BS5
H
E1H
BS109-1
100bp
109-1RH

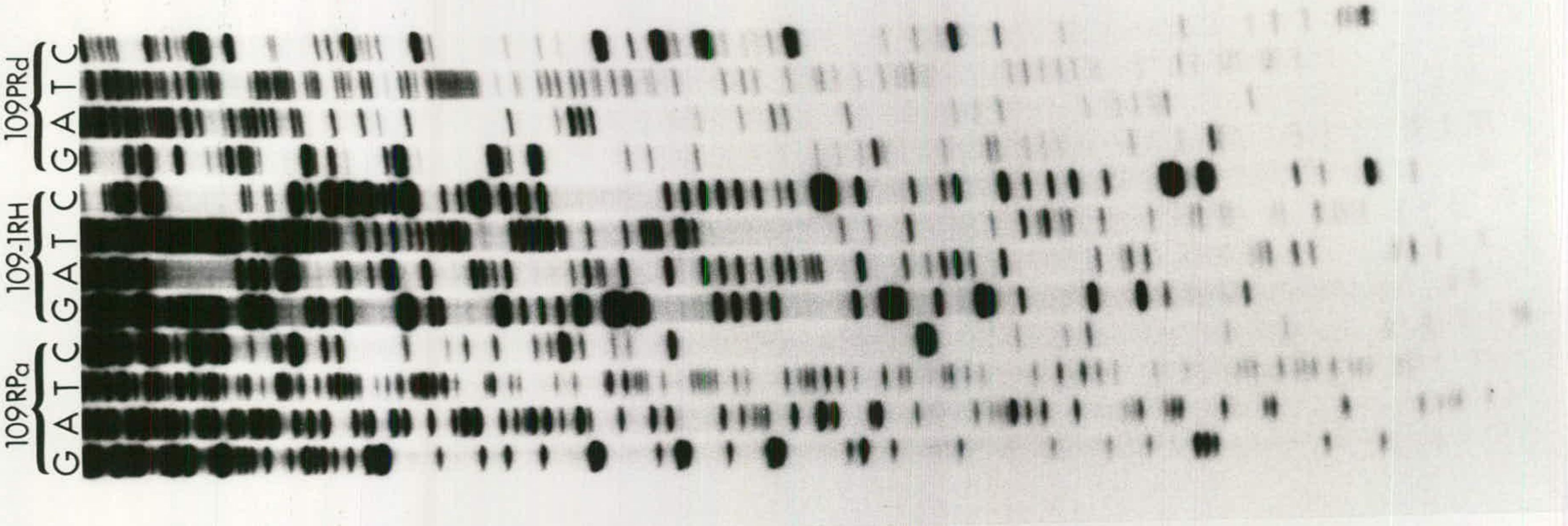
the problems associated with the Hong method. Furthermore, the breakpoints within the insert can be controlled to some extent and mostly cluster within about 150-bp of the target.

In conclusion, the chain terminator method is mainly chosen for large scale sequencing excercises because of the fairly large repertoire of procedures available for generating and screening primary and secondary M13 recombinants. However, every technique has its drawbacks. An example in the present case is the inability of the polymerase to 'read through' certain regions of secondary structure. Although this problem may be overcome by using higher incubation temperatures during synthesis or by using reverse transcriptase, other problems related to the cloning steps sometimes arise. An example of this is a failure to clone a particular DNA sequence either in one or in both orientations, due to the ability of the insert to express beta-galactosidase $\bar{A}$-donor activity, thereby producing only blue plaques (Close et al, 1983). In such uncommon situations it may be necessary to determine the sequence by the Maxam and Gilbert method.

FIGURE 5

A Representive Example of a sequencing Gel.

Sequencing reactions of three M13 clones (109RPa, 109-1RH and 109PRd) were run on a non-gradient short run (2.5 hours) 6% Acrylamide 8M Urea sequencing gel. The chain-terminators are indicated by G,A,T,C and the clones to which the sequences correspond to are indicated above the tracks.


109RPa
G A T C
109-1RH
G A T C
109PRd
G A T C

COMPARISON OF GROUP 2 GENES

Sequence data from about -1050 to +630 was collated from three Group 2 genes (BS109-2, BS102-2 and BL25) and compared to BS2, a Group 2 gene whose complete sequence is known (Clark *et al*, 1985a). This is shown in Figure (6) together with a table showing the frequencies of nucleotide differences between the genes. The Group 2 genes are predomonantly organised in a head to head linkage with Group 1 genes

Transcriptional Signals

The cap site of the Group 2 genes is putatively placed 31-bp down stream from the TATA box. This position is identical to the Group 1 start site (see chapter 2). The TATA boxes of BL25 and BS102-2 are identical to the Group 1 genes. However, BS2 and BS109-2 have point mutations around this region (Table 2). A comparison of the MUP gene TATA boxes with the consensus sequence TATAWAW drawn from 60 eukaryotic genes (Breathnach & Chambon, 1981; Table 2) shows that the G residue in the ninth position of BS2 is absent from all the other genes while the T in the first position of BS109-2 is present in only ten of the genes. Point mutations within the TATA box have been found to result in a marked reduction in transcription efficiency (Dierks *et al*, 1983). Therefore, the transversion events of BS2 and BS109-2 may have a down regulatory effect on their transcriptional efficiency. Interestingly, hybridization of a Group 2 probe under stringent conditions shows that Group 2 'like' transcripts are present in the low abundance short MUP mRNA of the male mouse liver (Clark et al, 1984a). The fact that Group 2 genes may be transcribed and correctly processed is also supported by the presence of a TATA box in these genes together with all the correct splice-donor acceptor sites of BS2 (Clark et al, 1985a; see Appendix).

Translational Signals

The initiation of translation in eukaryotic mRNA usually occurs at the first AUG codon that lies within a favourable sequence environment (Kozak, 1984a). A comparison of 211 mRNA leader sequences has revealed a conserved sequence immediately upstream of the AUG of the N-terminal methione. The consensus of this sequence is CCRCC where the R, usually an A at -3, is the most highly conserved residue. Mutations at -3 have been shown to reduce the efficiency of translation (Kozak, 1984b). The sequence ACCAA immediatly 5' to AUG in BS2, BS109-2, and BS102-2 has only one nucleotide in common with the consensus. Moreover the C at position -3 raises doubts as to whether these transcripts would effectively initiate translation. Note that BL25 retains an A at positon -3 but has only this in common with the consensus sequence.

Other translational differences seen are the inframe increase in the length of the signal peptide. The shortest is BL25 with 19 amino acids and the longest with 25 amino acids of BS102-2. The sequence in this region is simple, suggesting that they may have been created by 'slippage' during DNA synthesis or repair (Ghosal &

Saedler, 1978) or by unequal crossing-over. Finally, the most important observation is that all 4 group 2 genes share a common stop codon in the first exon (see Ghazal *et al*, 1985 and also the underlining in figure 6), suggesting that they have arisen from a common ancestral pseudogene. Other genetic lesions are seen in BL25 (Ghazal *et al*, 1985) and in BS2 (Clark *et al*, 1985a; see Appendix). This data has been published and is presented as part of this chapter. A possible function for the truncated product of the Group 2 genes is also postulated in Clark *et al* (1985a) (see Appendix).

Evolutionary Implications

All of the Group 2 sequences are closley related. The average divergence from the Group 2 consensus sequence is 1.4%. However, BL25 shows the highest divergence of 3.2 % while BS2, BS102-2 and BS109-2 diverge by an average of 0.7 %. The greater divergence of BL25 may be explained by the model of MUP gene evolution presented in Ghazal *et al* (1985). This model proposes that rather than the individual genes the 45kb unit containing the head to head linkage of a Group 1 and a Group 2 gene is the major unit of organisation and evolution of the MUP gene family. It permits either or both of the two evolutionary mechanisms, unequal crossing over and gene conversion, which are proposed to homogenize the gene family. Thus BS2, BS102-2 and BS109-2 may be derived from a common ancestor by more recent duplications or replacements and may be related to BL25 by more remote events.

The nucleotide differences between the Group 2 genes are

FIGURE 6.

Sequence Comparison of Group 2 Genes.

The 5' sequences of the Group 2 genes were aligned to maximize homology using the GAP programme of Devereaux et al (1984).
The numbers represent distances, in bp, from the cap site.
***, indicates the mapped position of nuclease hypersensitive domains (J. Clark unpublished).
N, is an undetermined base.
The TATA box and stop codons are underlined.

The Table lists the frequency of base changes between the genes.

```
    -1051                                                               -1002
BS2                                            A               G
BS102-2               NNNN NNNN                C
BS109-2               NNNN NNNN                C                  T
BL25                          C              C A
Consensus       gaattcttta tattcccaca tcaaacatt- gtttatgttt ctaattccag

    -1001                                                               -952
BS2                                              G
BS102-2
BS109-2
BL25
Consensus       ggtaaatgaa atgtctactg atacaaaata tgttcccatt gtaagtgtat

    -951                                                                -902
BS2             G   T               G G              G
BS102-2         A   A
BS109-2         A   A
BL25            GA  T                                        C
Consensus       -ttt-tggaa atatgatttt tggcatgatt tctacttgga catccatcag
                ********** ********** ********** ********** **********

    -901                                                                -852
BS2
BS102-2
BS109-2
BL25                    C                                   A
Consensus       caaatatgac tttgaacaat acttgtattt ttcattaatg gtgagattat
                ********** ********** ********** ********** **********

    -851                                                                -802
BS2               GTT          G
BS102-2            TT
BS109-2            AA
BL25               AA      G
Consensus       gtc--gtcag ttgaagtcca atgttctcaa aagtaaacag ttgatgaaac

    -801                                                                -752
BS2                              G
BS102-2
BS109-2                                                     C
BL25                        C                          A
Consensus       ctggctactg ataccacttt ggattctgag gtcagtgcta tctgtacaaa

    -751                                                                -702
BS2
BS102-2
BS109-2
BL25                              T C
Consensus       ggttgaagcc tttttggcag ctatccccat agccttggaa gacttttctt

    -701                                                                -652
BS2                 C
BS102-2             A         G
BS109-2             C
BL25                A
Consensus       gata-gtttt ttaagttaac atttactgtt tttatgtgta tgtgcctgaa

    -651                                                                -602
BS2                   T                     R  N
BS102-2               C
BS109-2               C
BL25                  T                                   N NNNNNNNNNN
Consensus       tgagtt-ata tgcaccactt gtgtacagga ggcaacaggg atcaaaagaa
                ********** ********** ********** ********** **********

    -601                                                                -552
BS2                          T
BS102-2
BS109-2                                                      A
BL25            NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN
Consensus       agtgtatggt ttcctgttac tgaagataga gtcagtttta ggatgcccag
                ********** ********** ********** ********** **********

    -551                                                                -502
BS2
BS102-2
BS109-2
BL25            NNN
Consensus       gtaagtgagg taattgaacc tgtgtcctca gcaaaaacaa gtttcaaaga

    -501                                                                -452
BS2                                                 C
BS102-2                                          Y
BS109-2
BL25                                         ACC C
Consensus       atctgattcc ttctttgacc tccactgcat taggcctgtg catggtactc

    -451                                                                -402
BS2                     A
BS102-2                 A
BS109-2                 T
BL25                    T   A
Consensus       aggtacac-t gcaaaaccaa tgctcataca catgaaatat gaatgaatct

    -401                                                                -352
BS2                G                                              A
BS102-2            T                                              C
BS109-2            T                                              A
BL25               G     A                                        C
Consensus       ttt-caaaag ttgagtaact caacttcttc catactcacc ttaaag-aaa

    -351                                                                -302
BS2
BS102-2
BS109-2
BL25                    CG                         C    C
Consensus       tattctatac atacttgtcc ttccagttca tttcccatcc cctccaaaac

    -301                                                                -252
BS2                            A
BS102-2
BS109-2
BL25
Consensus       tgtctttctg attccaagcc agatccaaag gtgcctctag ttccagatca

    -251                                                                -202
BS2
BS102-2
BS109-2
BL25
Consensus       cagttccctt gaacacccac tgttttcttg ggaatatgtt ttgagaaatg

    -201                                                                -152
BS2
BS102-2
BS109-2
BL25             G
Consensus       taagactact aaatcaatcc ataggtgatg atattgccaa gtttgaaaat
```

```
    -151                                                                -103
BS2
BS102-2
BS109-2                                        T
BL25                         T
Consensus       ggcaggaaca atccttggcc tctcatcaat aaatgagaaa atattccac

    -102                                                                -53
BS2                                                                  A
BS102-2
BS109-2
BL25                                   A            A           T
Consensus       aaagcctgac agaggtagag tcgacccata caggaagaaa aaaaaaaaa.

    -52                                                                 -3
BS2             AAA                         G
BS102-2
BS109-2                            T                     G
BL25
Consensus       ...cccactg aacccagaga gtatataagg acaagcaaag gagctgggga

    -2                                                                  +48
BS2               |Exon 1
BS102-2           |Leader Sequence
BS109-2           |
BL25              |G                              T
Consensus       gt|agagtgta ggcaacatca ccagaaagac gtggtcctga cacacagata

    +49                                                                 +98
BS2                                |Signal    ... CTGCTG
BS102-2                            |Peptide   CAG CTGCTG
BS109-2                            |          CAG ......
BL25                          AA  A|      A C.... .......... ....A
Consensus       attctatttc agaccaa.|at gaagcag--- ------ctgc tgctgctgct

    +99                                                                 +148
BS2                                                            |Mature
BS102-2            C                                           |MUP
BS109-2                                      T    C            |
BL25                G             A      T                     |   T
Consensus       gctgctgctg tgtttgggac taaccctagt ctgtgtccat gca|gaagaag

    +149                                                                +198
BS2                                                   |Intron I
BS102-2                            C                  |
BS109-2                                               |
BL25                                                  |
Consensus       ctaggtctat gtgaaggaac tttaatatag aaaaa|gtacg atcagtgaat

    +199                                                                +248
BS2
BS102-2
BS109-2
BL25                                        GAA                     A
Consensus       tgtatgtatg atcagaatgt gctttgtgga aatgttttag ccaagtgggt

    +249                                                                +298
BS2                                                      T
BS102-2
BS109-2
BL25              ........ .......... .......... .......... ..........
Consensus       cctttgaggg aatggttatt gtgccacaat gtattagaca aatgaatggc

    +299                                                                +348
BS2                    T                                           A
BS102-2                               A
BS109-2
BL25            .......... .......... .......... .......... ..........
Consensus       tccagacctt gaagtaagac cagctactca gatctaacaa tgtttgggga

    +349                                                                +398
BS2                                           T
BS102-2
BS109-2                                                  AC
BL25            .......... .......... .......... .......... ..........
Consensus       cctctgttct cactgagagg tagggacagt atgctaagca ttgtggcaac

    +399                                                                +448
BS2                     G
BS102-2                                          A
BS109-2
BL25            .......... .......... .......... .......... ..........
Consensus       tggcatgaaa tatacccctg tgtatgctca ggccctaatg cataggctgg

    +449                                                                +498
BS2                                      T
BS102-2
BS109-2
BL25            .......... .......... .......... .......... ..........
Consensus       agttcaagtg gaaacatgca tgtggtatcc ctgcttcttc ttccttaacc

    +499                                                                +548
BS2                           C     |Exon 2
BS102-2                             |
BS109-2                             |
BL25            .......... .........|. ......... .......... ..........
Consensus       atttgcttat ctagtacag|a ttaatggga atggtacact attatcctgg

    +549                                                                +598
BS2
BS102-2
BS109-2
BL25            .......... .......... .......... .......... ..........
Consensus       cttctgacaa aagagaaaag atagaagaac atggcagcat gagactttt

    +599                                                                +627
BS2
BS102-2                             N NNN
BS109-2          A  NNNNNN NNNNNNNNNN NNN
BL25            .......... .......... ...
Consensus       gtggagcaca ttcatgtctt ggagaattc
```

FREQUENCY OF BASE PAIR CHANGES BETWEEN GROUP 2 GENES.

	BS2	BS102-2	BS109-2	BL25
BS2		1.9%	2.5%	4.7%
BS102-2			1.3%	4.1%
BS109-2				4.3%
BL25				
CONSENSUS	1.0%	0.5%	0.7%	3.2%

randomly distributed over the sequences that can be compared. At 10 different sites within this sequenced region two different nuleotides are shared between the four genes. The pair of genes which have the same nucleotides are however different at the different sites. Therefore, this makes the construction of a simple pedigree relationship between the genes impossible. It is unlikely that such a distribution of nucleotides is due to independent random mutational events. In general, such observations are in agreement with the suggestion that Group 2 genes are related to each other through recent unequal crossing over or conversion events.

COMPARISON OF GROUP 1 GENES

Sequence data from about -350 to about +280 was collated from four Group 1 genes (BS1, BL1, BS5 and BS109-1) and compared to BS6, a Group 1 gene whose complete nucleotide sequence has been determined (Clark et al 1985a; see Appendix). This is shown in Figure 7. These genes are predominantly organised in a head to head linkage with Group 2 genes.

Transcriptional Signals

The cap site of the Group 1 genes is positioned 31-bp downstream from the consensus TATA box as previously determined. The TATA box region of all five genes is identical except for BS109-1 which has a T in position 3 of the consensus (see Table 3). This residue is a T in only 10% of the 60 genes compared by Breathnach and Chambon (1981). Whether such a difference has an effect on transcriptional efficiency is not known. However, in general the Group 1 genes have a very good fit to the consensus TATA box. Further aspects of the promoter region are discussed in Chapter 4. Nevertheless, notice the variable A-rich region situated at about -50 and the asymmetric distribution of base changes. The divergences seen at the 5' end of these genes approach a similar frequency as those between the Group 2 genes (Table 4). Such differences are speculated to be highly significant in consideration of the marked differences in expression of various MUP genes between and within secretory tissues.

FIGURE 7.

Sequence Comparison of Group 1 Genes.

The 5' sequences of the Group 1 genes were aligned using the GAP programme of Devereaux et al (1984).

The numbers refer to distances, in bp, from the cap site and the dashed lines represent undetermined sequences. Dots represent the gaps introduced to maximize homolgy. The TATA box is underlined.

```
    -679
BS1                                                                                                       C
BS6           ---------- ---------- ---------- ---------- ----------  ---------- ---------- ---------- --- G
Consensus     ggccacaagg gtcgaaagta tgtgatttcc tgttcctgat gttagagtgt  gttttgggat gccatgggag tgctggtaat tgaa-ctggg ttctcagtaa

    -579
BS1            T                             T                                            G
BS6            A                             A                                            A
Consensus     a-acaagttt caaagaatct gactccttc- tctgacctcc gctgcaccag  gcctggacat ggtactcag- tacacatgaa aaaccagtgc tcatacacat

    -479
BS1                  A                                                                           G
BS6                  G                                                                           A
Consensus     gaaatat-aa tgaatctttt gcaaaaattg ggtaactcaa gttcttccat  actcatctta aagcaaagat tctat-cata cttgtccttc cagttcagtc

    -379
BS1                                     T
BS6                                     C                   C
BL1           ---------- ---------- ------
BS109-1       ---------- ---------- ---------- ---------- ----------  ---------- ----------
Consensus     tgtactccct ccaaaactgg cttt-tgatt ccaaaccaga tcgaaagttg  catctggttc cagttggcag ttctcttgaa cacccactgt tttattggga

    -279
BS1                                                                                                           NN
BS6                        T
BL1
BS109-1                                                         C                  T
BS5           ---------- ---------- ---------- ---------- ----------  ---------- ---------- ---------- ----
Consensus     atatgttttg agaaacataa gattactaaa tcaatccata ggttatgaaa  ttgccaagtt tgcaaagggc aaggaacaat tcttggcctc taatcaataa

    -179
BS1                                                                        AT                AAAAAAA AAAAAAAAAA AACAAACAAA
BS6                                                                                       C
BL1                                     A                  T                                   AAA
BS109-1                                    T                                A C        C . ...
BS5                                 C                                                    C ...
Consensus     atgagaaaac attccacaaa gcctgacaga ggtagaggag acccatacgg  gaagagggaa aaaaaaaaaa aaa....... .......... ..........

    -79
BS1           C      A   A  A  A  A               A                                                  | Exon 1
BS6                               A .                                                                | Leader sequence
BL1           .......... .......... .......... ..                                                    |
BS109-1       .C                             T A          T      T                                   |               G
BS5              CA         G                           A                                            |
Consensus     aaaacaacaa caacaacaac aaaaaaaaaa cccgctgaac ccagagagta  tataaggaca agcaaagggg ctggggagt|g gagtgtagcc acgatcacaa

    +22
BS1                                                              | Signal                                            |
BS6                                                              | Peptide                                           |
BL1                                                              |                                                   |
BS109-1             T                                            |                                                   |
BS5                                                              |                                                   |
Consensus     gaaagacgtg gtcctgacag acagacaatc ctattcccta ccaaa|atgaa gatgctgctg ctgctgtgtt tgggactgac cctagtctgt gtccatgca|g

    +122
BS1           Mature                                       | Intron I
BS6           MUP Protein                                  |
BL1                                                        |                     ---------- ---------- ---------- ----------
BS109-1                                                    |
BS5                                                        |
Consensus     aagaagctag ttctacggga aggaacttta atgtagaaaa g|gtatgatca ctgaatagta gcttctgact cagaatgtgc tttggggaac tcttgaagcc

    +222
BS1
BS6
BL1
BS5
Consensus     aagtaggtcc tttgagggga tgggtatagt gccccaatct cttagacaaa  tgaatggatc c
```

Translational Signals

Several features which distinguish the Group 1 genes from the Group 2 genes suggest that they are the functional members of the family: 1) The sequence CCAAA found immediately 5' of the Group 1 initiation codon shows good homology with the consensus CCRÇC. 2) In sharp contrast to the Group 2 genes, the five Group 1 genes are 100% homologous within the first exon as well as part of the first intron, except for BS109-1 which has two point mutations within the leader sequence. 3) The length of the Group 1 gene signal peptide region is conserved. 4) The exonic sequences of BS1, BS6, BL1 and BS5 have been determined (Clark et al, 1985b). These data show a low frequency of base differences between the genes (within this region), on average about 0.3%, and most importantly that they code for full-length MUP polypeptides.

Thus, it is argued that such features reflect selective constraints imposed on the Group 1 genes, in agreement with the suggestion that the Group 1 genes are the functional members of the MUP gene family.

Evolutionary Implications

The frequency of base differences between the Group 1 genes are nonrandomly distributed over the sequence lengths that can be compared (Figure 7 and Table 4). As discussed previously the transcribed regions of these genes are highly conserved. However sequences upstream of the cap site, although markedly similar, have

TABLE 4.

Shows the frequency of nucleotide changes between the Group 1 genes.

Percentage differences observed between the total sequenced regions is shown on the right-hand side of the diagonal.
The differences between the sequences 5' to the cap site is shown on the left-hand side of the diagonal.

Frequency of base changes between the 5' sequences of the Group 1 genes.

			Total Sequenced Region				
	Gene		BS6	BS1	BL1	BS5	BS109-1
		Length (bp)	850	961	579	447	445
	BS6	568	-	1.3%	0.6%	0.4%	3.0%
Sequence	BS1	679	3.0%	-	0.6%	1.0%	4.5%
Upstream	BL1	297	1.7%	1.7%	-	0.7%	2.4%
of the	BS5	165	1.9%	5.0%	3.2%	-	3.5%
Cap Site	BS109-1	280	4.2%	6.0%	3.6%	6.3%	-

a higher frequency of base differences. This suggests that the Group 1 genes are related to each other through recent unequal crossing over or gene conversion and that such events have occured more frequently within the transcription unit. Since the Group 1 genes are the predominant functional members of the family such conservation of the 'coding' sequences may reflect selective constraints imposed on these genes.

FIGURE 8.

N-Terminal Sequences of BS6, MUP15 and BS105A.

The translation products of the first and part of the second exons of BS6, MUP15 and BS105A is shown. The letters correspond to the IUB code for the amino acids.

N-TERMINAL SEQUENCES OF BS6, MUP15 AND BS105A

```
                      Signal Peptide  | Mature MUP
       1                              |
BS6                 M       G         |     TG           E  H             DN NF   L L Q H  69
MUP15           LLLP              I   |  S  ME       Q S Y  FS AE  YE        S    A   N T
BS105A                                |     ERQ          F CK                T    V   H DN
Consensus     mk....llll lcleltlvcv ha|eeass--r nfnveking- w-tiilasdk rekieehg-m r-fve-i-v
```

COMPARISON OF BS105A WITH A GROUP 1 AND A GROUP 2 GENE.

Part sequence data was obtained of the lambda clone BS105A which contains a MUP gene that belongs neither to Group 1 nor to Group 2 and is therefore placed in a third, heterogenous group of MUP genes, called Group 3. The comparison of the 5' end of BS105A with BS6 (a Group 1 gene) and BS2 (a Group 2 gene) is shown in Figure (9); also shown is the frequency of base changes between these genes. BS105-A is convergently linked to a truncated Group 1 pseudogene. Its chromosomal environment is therefore different from the predominant MUP gene organisation.

Primary Structure

The sequences immediately 5' to the AUG initiation codon and the length of the signal peptide are identical to BS6. The coding region sequenced (exon 1 and part of exon 2) contain no deleterious mutations . The translational product of the mature MUP does not correspond to any known sequenced MUP protein or gene. A list of the major N-terminal sequences is shown in Figure (8). Note that although the region around the TATA box was not sequenced accurately, approximate sequence data can be gained from the top of a sequencing gel. This data suggest that BS105A has a TATA box and a short poly A stretch (about 15-20 As and 1-2 Cs).

FIGURE 9.

Sequence Comparison of BS105A with BS6 and BS2.

The sequences of the genes were aligned to maximize homology using the GAP programme of Devereaux _et al_ (1984).

The numbers refer to the distance, in bp, from the cap site. The stop codon and TATA box sequences are underlined.

The Tables refer to the frequency of base changes between the genes.

```
    -1086

BS2                                        T    T     T
BS105A              NNNN NN                C    C     C
Consensus     yaattcttta tattcccaca tcaaaca-ta g-ttatg-tt ctagttccag

    -1036

BS2                                            GGT C          G
BS105A                                         TAG T          C
Consensus     ggtaaatgaa atgtctactg atacaaaata ---t-ccatt gtaa-tgtat

    -986

BS2             T        A          GG  CAT         G      CA  C  C
BS105A          C        G C        TT  TGA         C      TC  G  T
Consensus     gt-tttggaa -t.atgattt --gg---gat ttcta-ttgg a--tc-at-a

    -936

BS2            C                                           G
BS105A         A                                           A
Consensus     g-aaatatga ctttgaacaa tacttgtatt tttcattaat g-tgagatta

    -886

BS2              GTT     G  G G       T          A  A  CA G   A  A
BS105A           CCA     T  C T       C          G  T  TT C   T  C
Consensus     tgt---gtca -tt-a-gtcc aa-gttctca aa-gt-aa-- -ttg-tg-aa

    -836

BS2                        T           A
BS105A                     C           T           NNNNNN NNNNNNNN
Consensus     cctggctact ga-accagtt tgg-ttctga ggtcagtgct atctgtacaa

    -786

BS2                      C    T       T               G          T
BS105A                   A    G       C               T          C
Consensus     aggttgaagc -tttt-ggca gc-atcccca tagcctt-ga agacttt-ct

    -736

BS2                CG         G     C       G                   C
BS105A             AA         A     G       A                   A
Consensus     tgata--ttt tttaa-ttaa -atttact-t ttttatgtgt atgtgc-tga

    -686

BS2                  T            T   T    R   N           A C
BS105A               C            A   A    A   A           G T
Consensus     atgagtt-at atgcaccac- tg-gtac-gg -ggcaacagg g-t-aaaaga

    -636

BS2                 A          G         G                AG      CA
BS105A              G          T         A                GA      AT
Consensus     aagtgt-tgg ttttct-tta ctgaa-atag agtcagtttt --gatgcc--

    -586
BS6           NNNNNNNNNN NNNNNNNN G       T      T
BS2            GT AGTGAG G   T         T
BS105A         TG GTGCTC A   A          A      G
Consensus     g--a------ -taa-tgaac ctgggtcctc agcaaaaaca agtttcaaag

    -536
BS6                          A           G                GA
BS2                  T                          TT  C
BS105A
Consensus     aatctgactc cttcttctga cctccactgc accaggcctg tgcatggtac

    -486
BS6               A         A        G
BS2                         C
BS105A                      T                         TT    A
Consensus     tcaggtacac atg-aaaacc aatgctcata cacatgaaat atgaatgaat

    -436
BS6                           G          G                T       C
BS2                       G                               C
BS105A        T      T              T           TT A      G
Consensus     cttttgcaaa aattgagtaa ctcaacttct tccatactca -cttaaagaa

    -386
BS6                                                 TGT   T
BS2             T                                T T    T
BS105A                C      GT   A
Consensus     aagattctat acatacttgt ccttccagtt cagtccccac cccctccaaa

    -336
BS6                               A                               T
BS2               T            A                 G   C    A
BS105A             A                 G          T  AT
Consensus     actggctttc tgattccaag ccagatccaa agttgcatct ggttccagat

    -286
BS6                                         A                    T
BS2           CA      C
BS105A             C
Consensus     ggcagttctc ttgaacaccc actgttttct tgggaatatg ttttgagaaa

    -236
BS6           CA     T                      T                    C
BS2                                               T
BS105A                         T      G
Consensus     tgtaagacta ctaaatcaat ccataggtga tgaaattgcc aagtttgaaa

    -186
BS6                                                   C
BS2            T              C           C                   T
BS105A                                                A
Consensus     agggcaagga acaattcttg gcctctaatc aataaatgag aaaa-attcc
```

```
    -136
BS6                                   G
BS2                                   TC          A       ..........
BS105A                                A CT           NNNN NNNNNNNNNN
Consensus     acaaagcctg acagaggtag ag-agaccca tacgggaaga gggaaaaaaa

    -86
BS6                                                       G
BS2           .......... .......... ..  A                 A
BS105A        .......... .......... ..NNNNNNNN NNNNNNNNNN NNNNNNNNNN
Consensus     aaaaaacaaa acaaacaaca acaacaaaaa aaaaaaaccc -ctgaaccca

    -36                                              | Exon 1
BS6                      A               G           |G         C  C
BS2                      G               A           |A         G  AC
BS105A        NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNN|NNNN NNNNNNNNN
Consensus     gagagtatat -aggacaagc aaagg-gctg gggagt|-gag tgtag-ca-g

    +15       Leader Sequence
BS6                                                  C
BS2                C                      C      T   T      T AG
BS105A              C  C G  GC    A     G          G G
Consensus     atcacaagaa agacgtggtc ctgacagaca gacaat-cta ttccctacca

    +65         |Signal
BS6             |Peptide                   A
BS2             |      CA GCTGCTGCTG CTGCTG
BS105A          |                             T     A              A
Consensus     aa|atgaag.. .......... ......ctgc tgctgctgct gtgtttggga

    +115
BS6                                     |Mature            C G
BS2             A                       |MUP         G     T T
BS105A             A            S       |                G A A  CA
Consensus     ctgaccctag tctgtgtcca tgca|gaagaa gctagttcta -g-gaaggaa

    +165                       | Intron I
BS6                            |          C      A
BS2                  A        A|   C                 .. GT ATGA
BS105A                         |
Consensus     ctttaatgta gaaaag|gtat gatcagtgaa ttgtagcttc tgactcagaa

    +215
BS6
BS2                    T     A G  T T        G               A    T
BS105A                    A          GT
Consensus     tgtgctttgg ggaactcttg aagccaagta ggtcctttga ggggatgggt

    +265
BS6                           C      C           A        C
BS2             T     A     G        T                    T
BS105A           T                   A             AG     A    T
Consensus     atagtgcccc aatctattag a-aaatgaat ggctccagac -ttgaagtaa
                             ****** ********** ********** **********

    +315
BS6                   C     T A  G  A      G            G
BS2                    C                       A       G
BS105A                                              C            T
Consensus     gaccagctat tcagatctaa caatgtttgg ggacctctat tctcactgag
              ********** ********** ********** ********** **********

    +365
BS6              C    G
BS2              T        T          C      G                 A
BS105A           A
Consensus     agg-agggac agtatgctaa gtattgtgac aactggcatg agatctaccc
              **********

    +415
BS6                                          A
BS2           C                      TG                       A
BS105A
Consensus     ttgtgtatgc tcaggcccta acccataggc tggagttcaa gtggacacat

    +465
BS6                                       C         G
BS2                   T                                           C
BS105A                                C
Consensus     gtgcatgtgg tatccctgct tcttcttcct taaccatttg cttatctatt

    +515
BS6               | Exon 2         C  T
BS2               |                T  C               T
BS105A            |         TTTT  TA  A
Consensus     acag|attaat ggggaatgg- a-actattat cctggcctct gacaaaagag

    +565
BS6                         TA       A T T          C        A
BS2                                  G                          T
BS105A                               C       G                   G
Consensus     aaaagataga agaacatggc a-catgagac tttttgtgga gcacatccat
                         ********** ********** **********

    +614
BS6
BS2
BS105A        CNNNNNNN
Consensus     gtcttggaga attc
```

A.

FREQUENCY OF BASE CHANGES BETWEEN THE GENES.

	BS2	BS105A
BS6	11.5%	10.8%
BS2		13.9%

B.

FREQUENCY OF BASE CHANGES WITHIN THE * LINED SEQUENCES

*** REGION	BS6	BS2	
+280 - +375	19%	11.5%	BS105A
+574 - +603	23%	6.6%	BS105A

Evolutionary Implications

BS105A appears to be equally diverged from BS2 and BS6, 13.9% and 10.8% respectively. The Group 2 genes have several characteristic sequence differences which distinguish them from the Group 1 genes. 1) A stop codon in the first exon; 2) a longer and variable signal peptide sequence; 3) a two base pair insertion/deletion followed by 6-bp of Group 2 specific sequence, located in intron I, 20-bp downstream from the intron boundary; 4) a two base pair insertion/deletion 53-bp upstream from exon 2; and 5) two 1-bp insertion /deletion events within the promoter region (positions -520 and -181 in Figure 9). With respect to these distinguishing features BS105A is identical to the Group 1 genes. This suggests that BS105A has probably evolved from an ancestral Group 1 gene and at some point has been converted/exchanged with a Group 1 gene. Although the base changes that have occurred between these genes appear to be random, short stretches of sequence about 20-bp in length are apparently more homologous with the Group 1 gene. These are indicated in Figure 9. Table B. This may represent a recent exchange with a Group 1 gene. However due to the divergences between these sequences and the shortness of these homologies, it is not possible to determine the significance of these observations. Nevertheless, it is proposed that during the evolution of BS105A exchange with Group 1 and Group 2 genes has taken place. This would be consistent with the model of MUP gene evolution which holds that the Group 1 and Group 2 genes are constrained from exchanging with one another by the fact that they are tightly linked in a 45-kb unit. Those MUP genes with a different unit of organisation, such as BS105A, may be released from such a constraint and are able to exchange with the different groups of MUP genes. This has been suggested for the MUP gene corresponding to MUP15 (Clark *et al*, 1985b).

CONCLUSIONS

Sequence data of the 5' end region of three Group 2 genes, four Group 1 genes and BS105A (a Group 3) gene has been determined. The sequence structure of these genes is consistent with the model of MUP gene evolution presented in Ghazal _et al_ (1985) and Clark _et al_ (1984b). The conservation of sequence that is observed within but not between the Groups of genes suggests that the 45-kb units of MUP gene organisation have evolved from a common ancestral 45-kb unit. Divergences between the Group 2 genes appear to have occured randomly, while those between the Group 1 genes have been nonrandom. The transcribed sequences of the Group 1 genes are more highly conserved than the 5' flanking sequences. This probably relates to selective constraints imposed on the genes. The greater divergences seen between the 5' flanking sequences of the Group 1 genes may reflect the differences seen in tissue-specific and hormonal regulation of these genes.

Finally, the most striking observation to be made from these data is that the Group 2 genes have most likely arisen from a common ancestral pseudogene. This suggests together with those truncated MUP pseudogenes that have been described, at least three (Clark _et al_, 1982; Clark _et al_, 1984b; Al-Shawi, 1985), that the total number of functional members of the family of about 35 genes is around 20. This number agrees well with the total number of different MUP proteins synthesised in the liver and lachrymal glands together.

STRUCTURAL CHARACTERIZATION OF THE MUP PROMOTER REGION.

INTRODUCTION

Viral enhancers and promoters have characteristic symmetries and repeated sequence motifs which are known to be important cis-regulatory elements. For example, the 72 and the 21 bp repeats of the SV40 Early Promoter region; and the HSV TK gene distal and proximal promoter elements both contain the sequence CCGCCC but in opposite orientations (McKnight et al, 1984).

Most if not all cellular Polymerase II genes which have been studied are also known to have multiple regulatory elements. For example, the promoter of the rabbit beta-globin gene contains a repeated 14 bp sequence, both copies of which are required for optimal transcription (Dierks et al, 1983); two copies of the heat shock regulatory site appear to be required for optimal induction in flies (Dudler and Travers, 1984); two steriod receptor binding sites have been mapped in the 5' flanking DNA of the chicken lysozyme gene (Renkawitz et al, 1984); the mouse mammary tumour virus long terminal repeats contain multiple binding sites for glucocorticoid receptors (Scheidereit et al, 1983), and a 12 base pair DNA motif which is necessary for metal regulation is repeated several times in the metallothionein gene promoters (Stuart et al, 1984). In most of the cases mentioned above, such sequence motifs are known to bind regulatory protein factors. The best studied examples of this are the 21-bp repeats of the SV40 early promotor and its binding of SP1 transcription factor; the Heat Shock Transcriptional Factor (HSTF) binding to the hsp70 gene (Parker and Topol, 1984; Wu, 1984); and the glucocorticiod hormone receptor complex and its binding to various viral and cellular promoters.

In general therefore, cellular Polymerase II promotors appear to be activated (or repressed?) by the multiple binding of trans-acting proteins/factors which recognize specific relatively short sequence motifs that are repeated several times within the 5' flanking regions of the genes. Thus, the analysis of the primary structure of a gene promotor may provide important clues as to the location and functionality of possible regulatory sequences. Futhermore, a search for those regulatory sequence motifs which have already been well characterized in other gene systems may provide insight into the regulation of the gene.

In this chapter I shall discuss the primary structure, the symmetries and those regulatory sequence motifs found in the 5' flanking sequences of the MUP genes.

A HYPERVARIABLE REGION IN THE MUP PROMOTER

A simple short repetitive sequence mainly of A residues is situated upstream of position -50 in the MUP genes and the alpha-2u globulin genes. This sequence varies strikingly in length and composition between 13 genes as is shown in Figure 10. The longest is the BS1 sequence with 73 nucleotides of almost pure A residues while the shortest is the BS109-2 sequence with only 11 As. On a closer examination the majority of the sequences may be divided into two types. 1) Those with the structure G_{n}---A_{11-18}---C_{n1}. These are the Group 2 genes, BL1, CL8, CL11, and alpha-2u globulin. 2) The other type has the structure G_{n}---$A_{n1}(CA_{2-3})_{n2}A_{n3}$---$C_{n4}$. These are BL7, BS5, BS109-1, BS6 and BS1. This is the major variability seen in the 5' flanking region of the MUP genes. Although the variability may have a trivial cause related to the repetitive nature of the sequence (see Chapter 3), the possibility that this region has a major functional significance is suggested by its close proximity to the TATA box. In this connection it is interesting to note the remarkable heterogeneity of the MUP gene family in tissue-specific expression and response to hormonal induction. Interestingly, three of the genes (BS6, BS109-1 and BL7) contain an enhancer core sequence within this region (CAAACaAC, see Table 5 site E-5). These three genes belong to Group 1 which contains the most abundantly transcribed MUP genes.

FIGURE 10.

A-TRACT REGION

Sequence alignment of the A-Tract region

Group 1 Genes:

BS1, BS6, BS109-1, BS5, BL7, BL1, CL8 and CL11.

(BL7,CL8 and CL11 were sequenced by R. Al-Shawi).

Group 2 Genes:

BS2, BS102-2, BS109-2 AND BL25.

Rat gene: alpha-2u-globulin.

A2u.91 (Kurtz, 1983).

```
                                                                                                                -31
BS1        GGGAAGAGG.ATAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAACAAACAAACAAAAAAAAAAAAAAAAAAAAAAAAAACCCACTGAACCCAGAGAGTATATAAGG
BS6        GGGAAGAGGG.AAAAAAAAAAAACAAAACAAACAACAACAACAAAAAAAAAAAA.............................CCCGCTGAACCCAGAGAGTATATAAGG
BS109-1    GGGAAGAGG...AACAAAAAACAACAACAACAACAACAACAACAAAAAAAATA...............................CCGCTGAACTCAGAGATTATATAAGG
BS5        GGGAAGAGGG.AAAAAAAAAACAAAACAAACAACAAGAACAACAAAAAAAAA...............................CCCGCTGAAACCAGAGAGTATATAAGG
BL7        GGGAAGAGGG.AAAAAAAAAAAAACAAAACAAACAACAACAACAAAAAAAA...............................CCCCGCTGAACCCAGAGAGTATATAAGG
BL1        GGGAAGAGGG.AAAAAAAAAAAAAAAAA.........................................................CGCTGAACCCAGAGAGTATATAAGG
CL8        GGGAAGAGGG.AAAAAAAAAAAAAAAA..........................................................CGCTGAACCCAGAGAGTATATAAGG
CL11       GGGAAGAGGG.AAAAAAAAAAAAAAA.........................................................CCCGCTGAACCCAGAGAGTATATAAGG
BS2        AGGAAG.....AAAAAAAAAAAAAAA.........................................................CCCACTGAACCCAGAGAGTATATGAGG
BS102-2    AGGAAG.....AAAAAAAAAAAA.............................................................CCACTGAACCCAGAGAGTATATAAGG
BS109-2    AGGAAG.....AAAAAAAAAAA.............................................................CCCACTGAACCCAGATAGTATATAAGG
BL25       AGAAAG.....AAAAAAATAAAA............................................................CCCACTGAACCCAGAGAGTATATAAGG
A2u.91     GAGAAGG....AAAAAAAAACAC................................................................CGAAACCCAGAGAGAGTATAAAG
```

Possible Functions For The A-Tract Region

1) Opening (Entry) Site for RNA Polymerase.

RNA Polymerase requires strand separation in order to synthesise mRNA (Saucier and Wang, 1972; Chamberlin, 1974). Since the hydrogen bonding of A:T base pairs is weaker than that of G:C base pairs, separation will be facilitated in A-rich regions, other things being equal. The melting of the strand would thus be inversely related to the length of the A-tract. In this connection it is interesting to note that those MUP genes with the longest A-tract belong to the highly expressed Group 1 genes.

2) Positioning of Regulatory Elements.

The effectiveness of upstream regulatory elements is often dependent on their distance from the TATA box (McKnight, 1982). However, the variability in length of this region is not very large in relation to the amount of lateral displacement that most upstream promoter elements can tolerate (up to about 50 to 100 bp Mcknight, 1982). It seems unlikely therefore that the function (if any) of the variable region is related to this phenomenon.

3) A Determinant For Nucleosome Phasing

Active genes are believed to be associated with regions (usually 5' and 3' to the gene) which are depleted of the regular nucleosomal coverage of chromatin (Elgin 1984). There is evidence to suggest that factors binding to the 5' flanking sequences prevent the formation (in vitro) of nucleosomes around this region (Emerson & Felsenfeld 1984). This suggests that the positioning of nucleosomes around the initiation site of transcription may well be an important regulatory mechanism. Although it is clear that there is a nonrandom component in the location of nucleosomes relative to the DNA sequence, both the extent of the nonrandomness and the mechanism(s) underlying the formation of 'phased' nucleosomes are controversial. If specific sequences or factors influence the position of nucleosomes then ideally these would be at or near to DNA sites required for the initiation of transcription. Recently it has been proposed that a protein factor (alpha-protein) purified from monkey cells mediates the positioning of nucleosomes (Strauss and Varshavsky, 1984). This protein is evolutionarily conserved and binds specifically to A, A+T rich (greater than 6-bp) double stranded DNA (Levinger and Varshavsky, 1982; Strauss and Varshavsky, 1984; Solomon et al, 1986). It is therefore possible that the A-tract region binds this alpha- protein. The proposed binding of this factor would be consistent with both positive and negative models of regulation of the MUP genes. According to the negative model the nucleosome would be so positioned as to prevent the binding of transcriptional factors. On the other hand the positive model proposes nucleosome positioning that would allow the interaction of transcriptional factors. Alternatively, although not exclusively distinct from the tissue-specific regulation, the length and composition of the A-tract region may determine the stoichiometry of binding such that this variability would relate to accessibility of the chromatin to various other transcriptonal factors. This would suggest that the different genes are turned more on or more off because of the sequences. More importantantly the essence of this model suggests the establishment of two distinct states along a short stretch of chromatin, one signaling 'on', the other 'off'.

REPEATS AND PALINDROMES

The direct repeats within the MUP Group 1 and Group 2 sequences were analysed using the UWGCG program REPEAT. This program identifies repeats by aligning two regions of the same sequence to reveal possible homology. In order to assess the significance of these repeats, two criteria were applied. Firstly, a direct repeat was considered significant only if two related sequences were not overlapping. This was to avoid including those repeats identified by the comparison of a sequence virtually with itself. Secondly, the direct repeat must be longer than that estimated to occur by chance alone, according to the statistics of Karlin et al (1983), including the standard deviation. This takes into consideration the length and nucleotide composition of the DNA sequence in question. From this analysis a direct repeat greater than 11-bp (9.4 ± 1.1) and 12-bp (10.8 ± 1.1) are significant for the Group 1 and Group 2 genes respectively.

The presence of inverted repeat sequences, potentially capable of forming stem and loop secondary structures was determined using the UWGCG program STEMLOOP. Stems over 12-bp in length and containing a maximum of 3 mismatches, were identified.

Figure (11) lists the symmetries found within the 5' flanking sequences of the MUP genes and Figure (12) shows the location of these on the BS1 and BS109-2 sequence. Symmetries are conserved between Group 1 genes and between Group 2 genes, but not between genes of the two groups.

GROUP 1 GENES

Direct Repeats

Two interrupted direct repeats of 14-bp and 15-bp are situated between nucleotides -500 and -380. These repeats are conserved in the Group 1 genes with 13 out of 14 matches between the two 14-mers (except for BS1: 12/14) and 3 mismatches for the 15-mer. The Group 2 genes show 11 matches between the two 14-mers and 9 between the two 15-mers. The upstream 14-bp repeats are displaced from one another by one helical turn. Thus the double-helix presents two identical steric patterns spaced apart by one turn of the helix. This may have biological significance if this repeat represents a protein binding site, since such a speculative interaction may involve binding a dimer of the protein or a monomer with a duplicated domain. There is no evidence that these repeats have any functional significance. However, in the rabbit beta-globin gene two 14-bp direct repeats are required for full expression (Dierks et al, 1983); also the human interferon alpha-1 gene contains four direct repeats of 6-8-bp within the upstream region responsible for viral induction (Ragg and Weissmann, 1983; Ryals et al, 1985).

Palindrome

A 26-bp palindrome, a feature characteristic of many protein recognition sites on DNA, is situated at about nucleotide -300. The Group 1 genes show 11 out of 12 matches between the arms of the palindrome except for BS1 and BL1 which show 10 out of 12. The Group

2 genes show 5 mismatches in the same regions. Palindromes have been observed upstream of a number of eukaryotic promoters and are concurrent with important regulatory elements. For example, alpha-1 globin genes have a 12-bp palindrome contained within a region known to be important for its expression (Mellon _et al_, 1981). The heat shock element (HSE) is an example of a short 10- bp palindrome (Pelham, 1982; Pelham and Bienz, 1982). In fact, a common feature of the heat shock promoters is a larger inverted repeat associated with the heat shock palindrome, although the sequence of the former is itself not conserved. These are the genes for hsp 70 (10 out of 12 bases form a dyad), hsp 83 (20 out of 24), hsp 22 (12 out of 16), hsp26 (12 out of 14), and hsp 27 (12 out of 16) (Holmgren _et al_, 1981). In the presence of the HSE there seems to be no correlation between the size or position of the larger inverted repeat and the efficiency of heat induction. Moreover, studies have shown that disruption of the large palindrome in the hsp 70 gene does not abolish activity (Pelham, 1982). These observations suggest that if there is any requirement for a palindrome , it need be no longer than 10-bp and that the sequence and exact position of the larger inverted repeats are unimportant. These studies involved the analysis of the _Drosophila_ promoter in COS cells and the requirement for the larger palindrome may be too subtle to be revealed by this assay system. The presence of such a feature in the same relative position in most of the _Drosophila_ heat shock genes certainly suggests that it has some function _in vivo_. The metallothionein promoters contain a 25-bp palindromic sequence around position -50 in the mouse MT-I and the human MT-IIA genes (Searle _et al_, 1984). It has been demonstrated that this

FIGURE 11.

SYMMETRIES WITHIN THE MUP PROMOTER REGION

Direct repeats and inverted repeats present within the Group 1 and Group 2 genes are shown. Subscript letters indicate those bases looped out from the alignment and the numbers refer to the position of these structures in the BS6 (Group1) and the BS2 (Group 2) genes.

Also shown is the mouse metallothionein I palindrome and its alignment with the Group 1 palindrome.

GROUP 1 GENES

DIRECT REPEATS

```
-482 ATACACATGAAAAA
     |||||||||||| |
-459 ATACACATGAAATA


-437 TCTTTTGCAAAAATT
     ||||  ||||| |||
-396 TCTTAAGCAAAGATT
          A
```

PALINDROME

```
-318 AACCAGATCCAAA
     |||||||| |||
-293 TTGGTCTACGTTG
```

GROUP 2 GENES

DIRECT REPEATS

```
-902 GCAAATATGATTTG
     ||||||||||||||
-859 GCAAATATGATTTG
              C

      G
-257 CAATCCAAAGGTG
     ||||||| |||||
-162 CAATCCATAGGTG
```

INVERTED REPEATS

```
-773 AAACAGTTGATG
     |||||||  |||
-627 TTTGTCATTTAC

-513 TGAACCTGTGTCC
     |||| ||||| ||
-420 ACATGGACTCAGG
               T
```

ALIGNMENT OF THE GROUP 1 PALINDROME WITH THE MOUSE METALLOTHIONEIN-I PALINDROME

MT-I PALINDROME

```
-67 CTGGGTGCAAAC
    | ||| ||||| C
-43 GGCCCGCGTTTC
```

```
GROUP 1   AACCAGATCCAAA GTTGCATCTGGTT
              |  | | ||||  ||||   | ||
MT-I          CTGGGTGCAAA TTTGCGCCCGG
                        CCC
```

represents a heavy-metal-inducing element (Stuart et al, 1984). However, disruption of the palindromic structure does not abolish metal induction, indicating that a palindrome itself is not essential for metal regulation (Carter et al, 1984; Searle et al, 1985). Interestingly, the core of the Group 1 palindrome shares sequence homology with the mouse MT-I gene palindrome and is shown in Figure 11. This may have either evolutionary or functional implications.

In conclusion, palindromic structures such as those so far described appear not to be essential, although they are associated with important regulatory elements. In this connection it is interesting to note that the Group 1 palindrome not only shares homology with the MT palindrome and the Metal Responsive Sequence Element but also with the enhancer core sequence and the Nuclear Factor 1 binding sites. These sites will be dealt with in the next section.

GROUP 2 GENES

Direct Repeats

The Group 2 genes have two interrupted direct repeats, a distal 14-bp and a proximal 13-bp repeat (Figure 11). The distal repeats are located within the nuclease hypersensitive domain present in both the Group 1 and the Group 2 genes. Whether these repeats are present in the Group 1 genes is not known since the sequence of this region has not been determined. The proximal repeats, only marginally represented in the Group 1 genes, flank the 5' end of two putative Nuclear Factor 1 binding sites (Figure 12). Additionally the upstream repeat is in the same relative position as the Group 1 palindrome and contains a putative Metal Responsive Element, while the downstream repeat contains homology with the Enhancer Core Sequence (Figure 12). The association of these repeats with various potential regulatory sequence motifs may suggest a functional significance.

Inverted Repeats

The Group 2 genes also have two interrupted palindromes or inverted repeats within their 5' flanking sequences (Figure 11). The upstream 12-bp inverted repeats flank the 3' and 5' side of the upstream and Group 2 specific hypersensitive domains respectively. Since the known Group 1 sequences do not extend this far, the presence of these symmetries within the Group 1 genes is not known. The downstream 13-bp inverted repeats are situated 3' to the Group 2 specific hypersensitive domain. The inverted repeat closest to this domain contains a Group 2 specific putative Glucocorticoid Responsive Element (GRE-1) (see Figure 12). The positon of these symmetries and their concurrence with known regulatory motifs suggests a possible significance for these sequences.

REGULATORY SEQUENCE MOTIFS?

A model of gene regulation holds that proteins in the cell nucleus function together as a network to mediate the extent of gene expression. Evidence for this model has been accumulating over the last five years in which the interactions of specific trans-acting proteins with cis-acting regulatory sequences have been identified. However, little is known about the structure and organisation of the control network as a whole. The total number of different transcription factors in the cell remains uncertain and the gross structural organisation of active genes within the nucleus is only begining to be resolved. In this section I shall describe those DNA-binding proteins that have been found to interact with polymerase II genes and their possible interactions with the MUP promoter sequences.

The UWGCG program FIND was used to search for regulatory sequence motifs within the Group 1 and the Group 2 genes. In order to determine the significance of these finds their expected and observed frequencies of occurence in the complete sequences of BS6 (4632-bp), BS2 (5023-bp) and pBR322 (4363-bp) were determined.

A summary of all known sequence motifs and their presence or absence within the MUP gene familiy 5' flanking sequences is shown in Table 5. Also, these sites are indicated on the aligned sequences of BS1 (Group 1 gene) and BS109-2 (Group-2 gene) in Figure 12.

From this survey, putative glucocorticoid responsive elements, enhancer core sequences, nuclear factor 1 binding sites and metal responsive elements were identified.

GLUCOCORTICOID CONSENSUS SEQUENCES

Glucocorticoids and other related steroid hormones (progesterone, oestrogen, and testosterone) modulate the transcription of specific genes in target cells by the interaction of the steroid-receptor complex directly with specific DNA sequences adjacent to the promoter of responsive genes. This is particularily well established for the glucocorticoid-receptor complex. Furthermore, these sequences appear to have enhancer like properties (Chandler et al, 1983). In all cases studied multiple sites are present. For example, the mouse MMTV has 4 sites (Scheidereit et al, 1983), the rabbit uteroglobulin gene has 3 sites (Cato et al, 1984), the chicken lysozyme and human MT-IIA genes have 2 sites each (Renkawitz et al, 1984; Karin et al, 1984). A computer search for homology in the MUP genes to the consensus sequences generated from 20 binding sites, SKKYWCMWYSTGTYCT (Dr. M. Beato, pers. communication), is shown in Table 5. Since the TGTNCT hexanucleotide is the most highly conserved feature of this consensus sequence only those finds containing this were chosen. Three putative sites are seen at the 5' end for each of the Group 1 and Group 2 genes. The BS2 gene (Group 2) contains an additional site at coordinate 3290, while no other sites are found in the BS6 (Group 1) gene or in pBR322. This would suggest that these finds within 600-bp of 5' flanking sequences are significant.

TABLE. 5

A SUMMARY OF REGULATORY SEQUENCE MOTIFS FOUND WITHIN THE MUP PROMOTER REGION

The approximate positions of these motifs refer to the BS6 (Group 1) and the BS2 (Group 2) sequences. 'rev') denotes the reverse sequence of the consensus and the lower case letters refer to mismatches with the consensus sequence. Subscripts below the motifs indicate those genes which differ from the sequence shown.

FACTOR/REGULATORY SEQUENCE	CONSENSUS	SITES	GROUP 1 GENES	APPROX POSITION	GROUP 2 GENES	APPROX POSITION
Nuclear Factor 1	TGGN$_{6-7}$KYCAM	a	TTGcAN$_7$CCA	-305	GTGcCN$_7$CCA	-270
(NF1)	rev)KTGRMN$_{6-7}$CCA	b			GTGAtN$_7$CCA	-225
		c	TGGN$_6$aTCAA	-168	TGGN$_6$aTCAA	-135
		d	TGGN$_6$GCCAC	-1		
		e			TGGN$_7$GCCAC	+240
		f	TTGAAN$_6$CCA	+285	TTGAAN$_6$CCA	+315
Metal Responsive Element	CYTTTGCRYYCG	1			aGGGatCAAAAG	-710
(MRE)	rev)CGRRYGCAAARG	2			aGG*ATGCccAGG A in BS107-2	-660
		3	CCgcTGCACCaG	-515		
	(Palindrome)	4	CaGATcC*AAAGt g in BS1, BL1	-315	CCaaTGC*A*T*T*aG accc in BL25	-480
		5	aGttTGC*AAAGG t in BS107-1	-195	CaGATcCAAAGG	-280
		6			CTAATGCAtAGg	+430
Enhancer Core Sequence	GTGGWWWG	1			CTATCCcC	-830
(E)	rev)CWWWCCAC	2	CAAACCAg	-320		
		3	CAATCCAt	-220	CAATCCAt	-185
		4	CATTCCAC	-140	tATTCCAC	-110
		5	CAAACaAC present only in BS4, BL7 & BS107-1	-75		
Glucocorticoid Responsive	SKKYWCMWYGTGTYCT	1			tTGaACc-TGTGTCCT	-535
Element	rev)AGRACASRWKCWRMMS	2	CTaTA*CAtACtTGTCCT ° in BS1	-380	CTaT*A*CAtACtTGTCCT co in BL25	-340
(GRE)		3	CaGTTCAgTCTGTaCT	-360		
		4	AGGACAaGcaaAGggG	-25	AGGACAaGcaaA*GgAG o in BL25	-25
CCAAT box	GGYCAATCT		GaCCcATaCa	-113	GaCCcATaCa	-81
Globin genes conserved -100 region 23-25bp upstream of CCAAT box	RRCYYCACCC		GAaaaCAttCC	-147	GAaaatAttCC	-105
Sp1	KGGGCGGRRY		NONE		NONE	
Heat Shock Element (HSE)	CTNGAANNTTCNAG		NONE		NONE	

Three of these sites GRE-1, GRE-2 and GRE-3 have good homology with the 16-bp consensus sequence, about 82%. In the highly expressed Group 1 genes GRE-2 and GRE-3, situated at about -380 and -360, are displaced from one another by one helical turn. The length of this displacement is a property characteristic of many of the Glucocorticoid Receptor binding sites, for example, Sites 2, 3 and 4 of the MMTV promoter (Scheidereit, et al, 1983) and the two sites present in the human MT-IIA gene (Karin et al, 1984). In these examples the proposed binding contacts of the receptor occur on the same face of the major groove of the DNA alpha helix. Such interactions may have a cooperative effect. In connection with this it is interesting to note that GRE-3 is totally absent in the Group 2 genes. However, the poorly expressed Group 2 genes while losing site 3 have gained GRE- 1, which is located 20-bp downstream from the group 2 specific DNase I hypersensitive domain. GRE-4 shared between Group 1 and the Group 2 genes is juxtaposed to the 3' of the TATA box and shows around 63% homology with the 16-bp consensus sequence.

Glucocorticoid regulation of MUPs has been reported for some strains of mice (Knopf et al, 1983). Interestingly, the progesterone receptor complex has been shown to bind to the same sequences as the glucocorticoid receptor complex on the chicken lysozyme gene (Renkawitz et al, 1984). This suggests that the GRE consensus sequence may be a general recognition sequence for steroid hormone receptor complexes. Therefore, the MUP GREs may be involved in regulation by other steroid hormones.

ENHANCER CORE SEQUENCES

Enhancers, cis-acting transcriptional control elements, have been described in both viral and cellular genes (Khoury and Gruss, 1983). They influence transcription in a quantitative fashion, act over relatively large distances (several kilobases) and behave independently of their position and orientation. Enhancers have been described in immunoglobulin, chymotrypsin and insulin genes (Mercola et al, 1983; Gillies et al, 1983; Queen and Baltimore, 1983; Picard and Schaffner, 1984; Banerji et al, 1983, Neuberger, 1983; Queen and Stafford, 1984; Walker et al, 1983. They bear little homology with each other except for an 8-bp 'consensus' core element, GTGGWWWG (Laimins et al, 1982 and Weiher et al, 1983), but even this element is sometimes non-homologous (Walker et al, 1983). I have searched for such an element in the MUP gene family. Allowing for a 1-bp mismatch 5 potential sites are seen within 600-bp of 5' flanking sequence. Only one of these in the Group 1 genes shows 100% homology to the consensus sequence. Counting the Ws as half sites, allowing for 1-bp mismatch and taking into account that the sequence may be represented on both strands, suggests on a random basis one enhancer core element every 205-bp. The observed frequency of this in the BS6, BS2 and pBR322 sequences is one per 211, 200 and 273-bp. Therefore, the two to three sequences found within the 5' flanking sequences may have no special significance. Whether these sequences do have any significance must await a functional assay. However, the analysis of the MUP Group 1 promoter region by transient expression in fibroblast cells has shown that while the promoter alone is inactive it can be activated by linking in cis the SV40 enhancer element (see chapter 5). This shows that the MUP promoter is responsive to enhancer sequences and suggests that the observed enhancer core motifs within the MUP promoter region, if they are functional, are part of a tissue-specific element that is regulated within this cell system.

NUCLEAR FACTOR 1 BINDING SITES

Nuclear Factor 1 is a eukaryotic nuclear protein that binds in vitro with high affinity to the specific nucleotide sequence, $TGGN_{6-7}KYCAM$ (Nagata et al, 1983; Rawlins et al, 1984; Siebenlist et al, 1984; Hennighausen et al, 1985, Borgmeyer et al, 1985). The protein was first identified as a component of a purified nuclear extract of uninfected HeLa cells able to support the in vitro replication of Adenovirus DNA upon binding to a specific sequence in the terminal repeat (Nagata et al, 1983; Rawlins et al, 1984). Nuclear Factor 1 activity has been detected in a wide variety of species (chicken, mouse, man and Drosophila) as well as in various tissues (B-lymphocytes, brain, kidney and liver) (Borgmeyer et al, 1984; Hennighausen et al, 1985; and Rawlins et al, 1984). This wide range of incidence and likely conservation suggests that the protein has a fundamental function. A function other than its involvement in DNA replication has been proposed because the factor is found to bind close to regulatory sequences (Nowock et al, 1985). For example, in the chicken lysozyme gene two binding sites have been found in the 5' flanking region. These coincide with a hypersensitive site found in active chromatin (Borgmeyer et al, 1984). Recently the same workers have found that the region contains

FIGURE 12

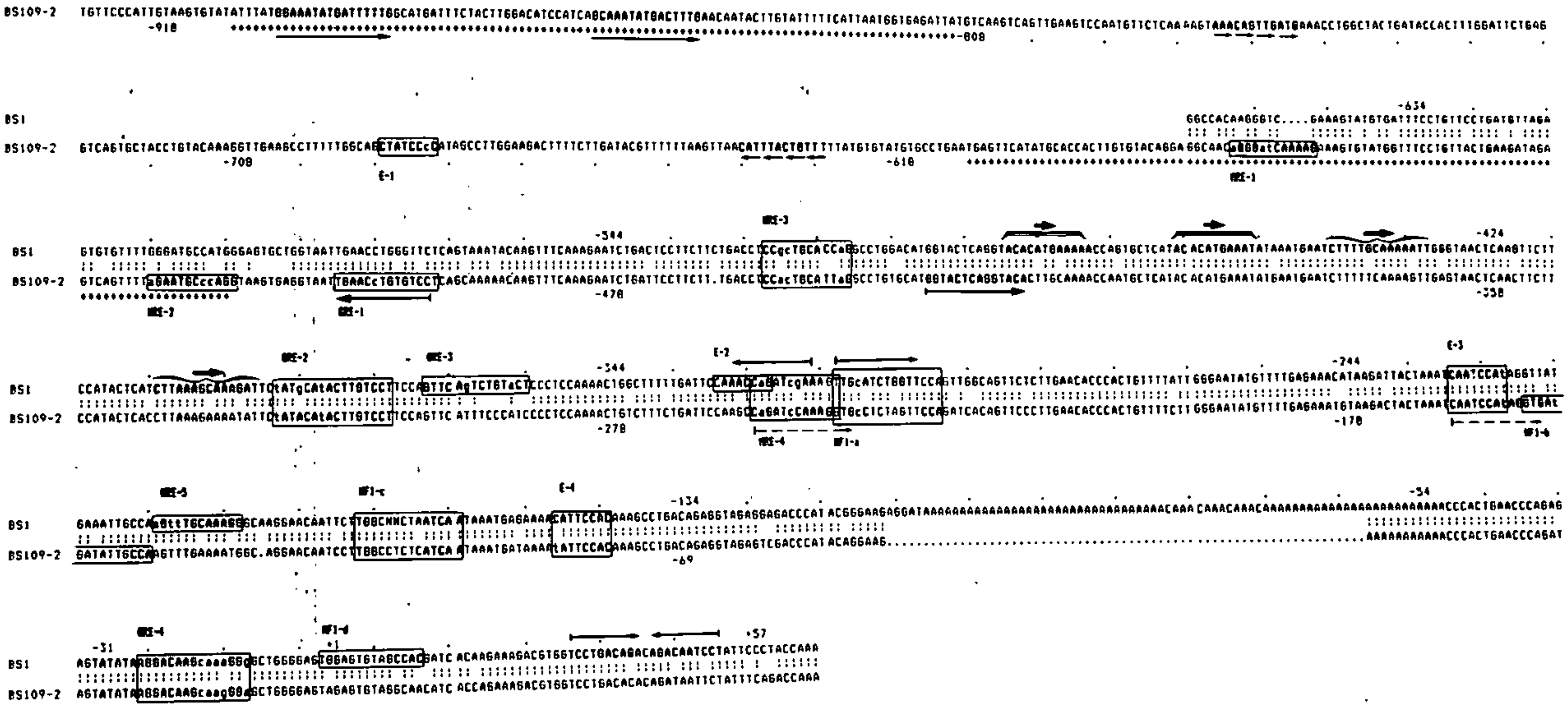
SEQUENCE STRUCTURE OF THE MUP PROMOTER REGION.

Sequence alignment of BS109-2 (Group 2) and BS1 (Group 1) genes is shown.

****, denotes the nuclease hypersensitive domains. The upstream domain is common to both the Group 1 and Group 2 genes while the downstream domain is present only in the Group 2 genes.

Repeats are indicated by arrows above the sequences for the Group 1 genes and below for the Group 2 genes.

Boxed sequences show the positions of regulatory motifs (see Table 5.) and the lower case letters mark those bases which differ from the consensus sequence.



BS109-2
-918
-808
BS1
-708
-618
-634
E-1
MRE-1
-544
-424
-478
-358
MRE-2
MRE-3
GRE-1
-344
-244
-278
-178
E-2
E-3
MRE-4
NF1-a
NF1-b
MRE-5
NF1-c
E-4
-134
-54
-69
-31
+1
+57
MRE-6
NF1-d

a tissue-specific enhancer (Arolla Workshop 1985). Two NF1 sites are found 3' to the IgM enhancer and two upstream of the human c-myc gene (Siebenlist et al, 1984). Additionally, recent observations in Dr. Hennighausen's laboratory suggest that NF1 is involved in transcriptional activation (Dr. L. Hennighausen pers. communication).

Allowing a 1-bp mismatch in the redundant sites of the consensus, three potential sites were found in the upstream region of each of the Group 1 and Group 2 genes (Table 5 and Figure 12). The expected frequency of occurence of the sequence on either strand of the DNA is one per 342-bp. The observed frequency in BS6, BS2 and pBR322 sequences is one every 260, 280 and 370-bp respectively. Therefore the frequency of its occurence in the 5' flanking sequences is not significantly different from random expectation. Furthermore, a cautionary note, which applies equally well to the other putative sequence motifs, is that even if a 100% fit to the consensus is seen it does not necessarily imply the binding of the factor. This is exemplified by the factors which bind to the chicken beta-globin gene. One or more partially purified protein factors isolated from adult chicken erythrocyte nuclei can interact specifically in vitro with the 5' end DNA fragments. These fragments contain the sequences that are nuclease hypersensitive in vivo (Emerson and Felsenfeld, 1984) and DNase footprinting reveals that the binding sites comprise two discrete regions, each about 25-bp long and separated by 15-bp. The 3' binding site contains a sequence that shares good homology to the NF1 binding site consensus sequence. However, NF1 has been shown conclusively not to be the protein involved in binding to this region (Emerson et al, 1985). This suggests that constraints other than the sequence motif are important in determining the specificity of binding. This point will be discussed later. Nevertheless, the fact that sequences containing homology with such consensus motifs occur within a potential control region of the genes suggests a possible significance. Confirmation of this would require binding studies as well as a functional assay system.

METAL RESPONSIVE ELEMENTS

A repeated sequence motif, CYTTTGCRYYCG, found within the 5' flanking sequences of the metallothionein genes is responsible for the heavy-metal regulation of these genes (Stuart et al, 1984). A computer search of the 5' flanking sequences of the MUP genes revealed reasonably good fits to this consensus sequence. A maximum of 3 mismatches were allowed for the 12-bp motif. The expected random frequency of occurrence on either strand is once every 108-bp and the observed frequency in BS6, BS2 and pBR322 is once every 260, 260 and 190-bp respectively. Therefore the frequency of these sites within the 5' flanking sequences of the MUP genes is again not significantly different from random expectation. However, one of the sites (MRE-4) lies within the Group 1 palindrome which shares homology with the mouse MT-I palindrome. The fact that multiple 'MRE-like' motifs are found and the homology to the MT-I palindrome strongly suggests the that MUP genes may be regulated by heavy metals. Heavy-metal regulation of MUPs has not been reported. However, if MUPs are controlled by heavy metals then they may have a function other than as pheromonal carriers or agents. Since MUPs are rapidly excreted into the urine a possible implication of these proteins in heavy-metal metabolism is very attractive. Testing the binding of heavy metals by MUPs and the induction of MUP transcription by heavy metals are relatively simple experiments which are currently in progress. Preliminary observations indicate that MUP mRNAs are indeed inducible by heavy-metals (J. Whitaker, unpublished) and implies that at least one of the MREs identified is functional.

MULTIPLE SEQUENCE MOTIFS

The structural organisation of many, if not all promoters recognised by RNA polymerase II involves multiple elements, usually located upstream of the start site, which are required for optimal and accurate initiation of transcription. These elements are in general relatively short nucleotide sequences of about 10-bp. Some of the elements occur only once and are position and orientation dependant, for example the TATA and CCAAT box sequences. Others, such as the Sp1 binding sites occur many times, perhaps necessarily, and are relatively position dependent but orientation independent. Also, there are regulatory elements which are both position and orientation independent. This class of regulatory elements have been termed enhancers. Within the MUP promoter region homologies to known regulatory sequence motifs are found to occur either once (these are the TATA box and a weak CCAAT box), or several times (these are the GREs, MREs and NF1 sites) while those motifs which are absent, Sp1 and HSE, are not even poorly represented. Although one may not expect to find homology with the HSE, the absence of Sp1 sites may reflect the tissue-specific nature of the MUP promoter. Those promoters that contain Sp1 recognition sites (SV40 early promoter, HSV TK gene and the metallothionein promoter; see SV40 early promoter in the introduction of Chapter 5) show little or no restriction of expression to particular cell types. Furthermore, housekeeping genes which contain a high GC content within their 5' flanking regions sush as the DHFR gene and the HMG Co-A gene contain multiple Sp1 sites (Reynolds _et al_, 1984; Dynan et al, 1986). This implies that the Sp1 transcription factor is not a tissue-specific factor and seems to be associated with 'housekeeping' functions.

One reason for the widespread multiplicity of regulatory sequences might be that duplication of the recognition sequence is a relatively simple evolutionary mechanism to increase the effect obtained with one copy. This could either be an additive effect if the regulatory proteins bind independently to the two sites, or a cooperative effect if binding of one regulatory protein facilitates binding of the second, perhaps by providing stabilising protein contacts or by creating a more favourable chromatin structure. Additionally, the presence of multiple elements might help keep a region of DNA accessible to RNA polymerase; thus, binding of a regulatory protein to one site might impose a particular phasing of nucleosomes, whereas binding to two appropiately spaced sites could exclude nucleosomes from the intervening DNA. The regulatory sequences that have been identified are relatively short and apparently fairly tolerant of nucleotide substitutions in several positions. Therefore, sequences able to bind the regulatory proteins would be expected to occur many times in the genome just by chance. Ensuring the specificity of transcriptional regulation must therefore involve other determinants or constraints. Such constraints might be imposed by a necessity for multiple regulatory proteins to be bound relatively closely together and in proximity to other promoter elements. Also, sequences around the recognition site may be important in determining the binding of a regulatory protein. In this case it is more likely the geometry of the DNA backbone that plays a role, since other specific sequences would probably have been identified. Sequence specificity on the basis of contacts between the DNA phosphate backbone and the protein surface has been termed 'structural' recognition, rather than 'sequence' recognition by Lomonossoff et al (1981).

In conclusion, the action of cis-acting regulatory sequences on gene expression may be determined by their interactions with one another, their binding of trans-acting factors, and the interactions between the bound trans-acting factors.

ENHANCER ACTIVATION OF THE MUP PROMOTER

INTRODUCTION

A large number of eukaryotic structural genes have been cloned and the detailed structural organisation and nucleotide sequences of many mammalian genes is known. Investigations of regulatory mechanisms have largely concentrated on defining the limits of the primary transcripts and identifying sequences involved in the initiation or promotion of transcription. These investigations have been greatly facilitated by the approaches of reverse genetics, that is, the introduction of cloned genes into cells or whole animals, and by the development of in vitro transcription systems. In the long term these, together with the purification of transcription factors, should lead to the elucidation of the molecular mechanisms underlying gene expression.

Expression of Genes in Tissue-Culture Cells

In general, the already large and rapidly expanding number of gene promoters that have been studied diplay a variety of native transcriptional activities. Such activity may be (1) highly tissue-specific, as is the case for the globin and MUP gene families; (2) widely-expressed, as in the case of genes associated with housekeeping functions, for example the DHFR and HMG CoA reductase genes; (3) widely-expressed at low levels but showing increased expression within particular tissues, as in the case of the metallothionein genes. Finally (4) some genes are induced by environmental or physiological signals, such as the metallothioneins genes to heavy-metal induction and the heat shock genes to temperature stress. A brief description of the expression of these various gene systems after transfection into tissue-culture cells will be given, followed by a discussion of the positive and negative regulation of eukaryotic genes. The SV40 early promoter region is used in the analysis of the MUP promoter region (in this chapter) and is also the most intensively studied of the eukaryotic promoter regions. An outline description of this region is also presented below.

Widely-Expressed Genes

Genes that are natively widely-expressed have been shown to be actively expressed following transfection into many different cell systems. These are the cellular housekeeping genes such as the HMG CoA reductase gene (Osborne et al, 1985), the mouse HPRT gene (Melton et al, 1986) and some viral promoters such as the HSV-Tk promoter (McKnight, 1982) and the SV40 early promoter. In some cases genes that are expressed in many different cell types are expressed more highly in a particular tissue. For example, the mouse metallothionein genes are maximally expressed in hepatic and renal tissues, while low levels of expression have been detected in many other tissues such as the spleen, intestine, heart, muscle, brain and testes (Durnam and Palmiter, 1981). Transfection of these genes into heterologous cell systems has shown that expression can be obtained but as yet throws no light on the tissue- specific nature of their expression (Searle et al, 1984).

Inducible Genes

There are many eukaryotic genes which are expressed in differentiated cells and which are rapidly activated or shut off in response to environmental changes. Such inducible gene systems have proven to be readily amenable to experimental analysis. For example, regulation of the expression of transfected genes by glucocorticoids has been achieved in three cases: MMTV (Chandler et al, 1983), the chicken lysozyme gene (Renkawitz et al, 1984) and the human metallothionein IIA gene (Karin et al, 1984). Other examples are heavy- metal regulation of the human and mouse metallothionein genes, viral induction of the alpha- and beta- interferons and the induction of the heat shock genes of drosophila (Pelham, 1982). In all of these studies cis- acting DNA sequences other than the TATA box and within 250-bp of the cap site were implicated in the regulation of gene transcription.

Tissue-Specific Genes

Tissue-specific expression of a number of genes has been demonstrated in transfected tissue-culture cells. That is, the correct expression of the gene has been shown to be restricted to a particular cell type. Examples are: the immunoglobulin genes (Foster et al, 1985), insulin and chymotrypsin genes (Episkopou et al, 1984; Walker et al, 1983) and the alpha-1 antitrypsin gene (Ciliberto et al, 1985). In all of these cases the 5' end region has been shown to be important in conferring (at least in part) the tissue-specific expression of the genes. In some cases (such as the immunoglobulin genes (Banerji et al, 1983; Gillies et al, 1983)) additional control sequences are found within the transcription unit. However, some highly tissue-specific genes may be expressed after transfection into heterologous cell systems. For example, short term experiments have shown that the alpha-globin gene promoter is functional at a low level upon transfection into fibroblast cells while the beta-globin gene promoter is essentially non-functional in such cells unless activated by, for example, a viral enhancer sequence linked in cis (Humphries et al, 1982; Banerji et al, 1981; Mellon et al, 1981).

The major limitation to the analysis of inducible and tissue-specific genes is that it is dependent on the availability of an appropriate cell line. A further complication is the fact that different immortilized cell lines from a given tissue show considerable variability in their phenotypic expression. For example no two hepatomas are exactly identical in their pattern of gene expression. In general, correct expression of a transfected gene is observed only in cell lines which express the corresponding endogenous gene (Ott et al, 1984). Additionally, such cell lines are always different from their tissue of origin in that they are immortal and therefore doubts will always be present as to the validity of the results obtained with them. In some cases such limitations may be overcome by introducing genes into primary cells. For example Renkawitz et al (1982) have shown hormonal inducibility of the chicken lysozyme gene after direct injection into chick primary oviduct cells. However, such experiments are difficult and very time consuming and further complicated by the fact that upon plating, changes in the pattern of gene expression may occur (Clayton and Darnell, 1983; see also 'MUP expression in hepatocytes' in the Introduction). In the long term transgenic animals will provide the solution to many of these limitations since from one animal it is possible to directly compare the tissue-specificity of expression of the transfected gene.

Positive Regulation

All of the abovementioned systems conform to a positive model of gene regulation. For example, those *trans*-acting factors that have been purified, Sp1, HSTF and steroid receptors, are positive transcription factors and have already been discussed in chapter 3. Indirect competition assays have also shown that positive *trans*-acting factors are involved in the expression of genes transfected into tissue-culture cells (Seguin *et al*, 1984; Scholer and Gruss, 1984).

Negative Regulation

There are several well documented examples of specific *trans*-acting repressors: a fibroblast chromosome that turns off specific liver functions (Killary and Fournier, 1984); the adenovirus E1a gene product, which represses the SV40 and polyoma enhancers (Borelli *et al*, 1984; Velcich and Ziff, 1984); the *trans*-acting negative regulation of viral enhancers (MSV) in undifferentiated embryonic stem cells (Gorman *et al*, 1985); and the T antigen repression of the SV40 early promoter (Tjian, 1981). In addition, Osborne *et al* (1985) have shown that 500-bp of sequence 5' to the HMG CoA reductase gene are necessary for constitutive expression in L cells as well as being responsible for cholesterol-mediated inhibition of transcription and suggest the possible involvement of sterol-binding repressor proteins. The increased expression of transiently expressed genes upon brief exposure to cycloheximide (an inhibitor of protein synthesis) also suggests the existence of a labile repressor(s) (Ishihara *et al*, 1984). In summary it appears that there are *cis*-acting sequences which respond to negative controlling factors and that they are located in the same, or in close proximity to, essential *cis*- acting sequences which mediate the positive control of gene expression.

The SV40 Early Promoter Region

Important *cis*-acting regulatory elements of the SV40 early promoter have been mapped, and reconstituted *in vitro* transcription reactions have allowed specific cellular factors that recognise and bind to the viral promoter to be identified and isolated.

The early genes of simian virus 40 (SV40) are expressed shortly after infection, whereas the late genes are maximally activated only after the onset of viral DNA replication and the repression of viral early transcription by T antigen (Tooze, 1980; Tjian, 1981). There are two T antigen binding sites surrounding the origin of replication and one that overlaps the first Sp1 binding site (see later). Mutational analysis of the viral transcriptional control sequences has revealed that the major early promoter consists of three 21-bp repeated elements preceded by a stretch of AT-rich sequences (TATA box), and early transcription has been shown to initiate from distinct sites located 20 to 30-bp downstream from the AT-rich region (Myers et al, 1981; Benoist and Chambon, 1981; Ghosh et al, 1981). In addition, enhancer elements that stimulate SV40 early transcription in vivo are located within the 72-bp repeated sequences, which lie 110 to 250-bp upstream from the early transcription start sites (Benoist and Chambon, 1981; Banerji et al, 1981; Gruss et al, 1981). These elements have been shown to interact with specific trans-acting factors (Scholer and Gruss, 1984) and increase the rate of transcription initiation by RNA polymerase II within linked sequences (Treisman and Maniatis, 1985; Weber and Schaffner, 1985). Fractionation of crude HeLa cell extracts resulted in the identification of a transcription factor, Sp1, that binds specifically to a hexanucleotide sequence, GGGCGG (GC-box), that is tandemly repeated six times in the 21-bp repeats of SV40 (Dynan and Tjian, 1983; Gidoni et al, 1984). Recently, Sp1 has been shown to activate transcription and bind to the GC-box sequences present in several other viral and cellular promoters, including the herpes virus IE-3, IE-4/5 and Tk promoters; the AIDS virus long terminal repeat (LTR); two monkey genomic promoters; and the human metallothionein gene promoters (Dynan et al, 1985; Jones and Tjian, 1985; Jones et al, 1985). Gidoni et al (1985) have shown that the first three proximal GC-boxes (I, II, and III) are involved in SV40 early RNA synthesis while Sp1 binding to sites III, V, and VI mediate late gene transcription. In conclusion, the approximately 300-bp of the SV40 early promoter region is a mosaic of different regulatory elements, which interact with different trans-acting proteins, and which sometimes overlap.

As discussed previously a large number of promoters work very weakly or not at all in some cell lines. However their activity can be forced to higher levels by linking enhancer sequences to them in cis. Examples are, the beta-globin, conalbumin and lysozyme promoters (Banerji et al, 1981; Wasylyk et al, 1983; Renkawitz et al, 1984). The SV40 enhancer, being one of the strongest enhancers and active in many cell types, is often chosen for this purpose. This chapter describes the activation of the MUP promoter by the SV40 enhancer element. Initially, J.O. Bishop had prepared constructs containing the MUP promoter with or without the SV40 early promoter region. Expression was observed only when the SV40 sequences where present. I shall describe in more detail the SV40 enhancer dependence of the tissue-specific MUP promoter.

PLASMID CONSTRUCTIONS AND NOMENCLATURE

A series of DNA constructs have been made in order to investigate the behaviour of the promoter region of one of the MUP genes. Each construction involves up to three components: (1) various 5' end regions of BS6 (a Group 1 gene), (2) the HSV thymidine kinase (Tk) gene as a reporter function, and (3) different regions of the SV40 early control sequences. The schematic arrangement of these structures is shown in Figure 13 and the specific sequence details of the regions are illustrated in Figures 14 and 15. A descriptive nomenclature identifies the various components associated with each vector. These various components are linked as shown into a section of pBR322 (coordinates 2069 to 4363) containing the plasmid origin of replication and the beta-lactamase gene for selection of the recombinant. The pSVEP. series of recombinants were constructed by Dr. J.O. Bishop, while I have prepared the pSVE. series by the in vitro manipulation of this series. The pSVEP. constructs contain the complete SV40 early promoter region which retains the enhancer element (the 72-bp repeats), the upstream promoter region (the 21-bp repeats and TATA box) and 61-bp of leader sequence, including the origin of replication. Digestion of this region with the restriction enzyme FokI cleaves the enhancer from the early promoter leaving 14-bp of the promoter (containing half of a 21-bp repeat). See Figure 15. This enhancer fragment (E) was cloned into the EcoRV site of M13tg130. The structure of the recombinant was confirmed by DNA

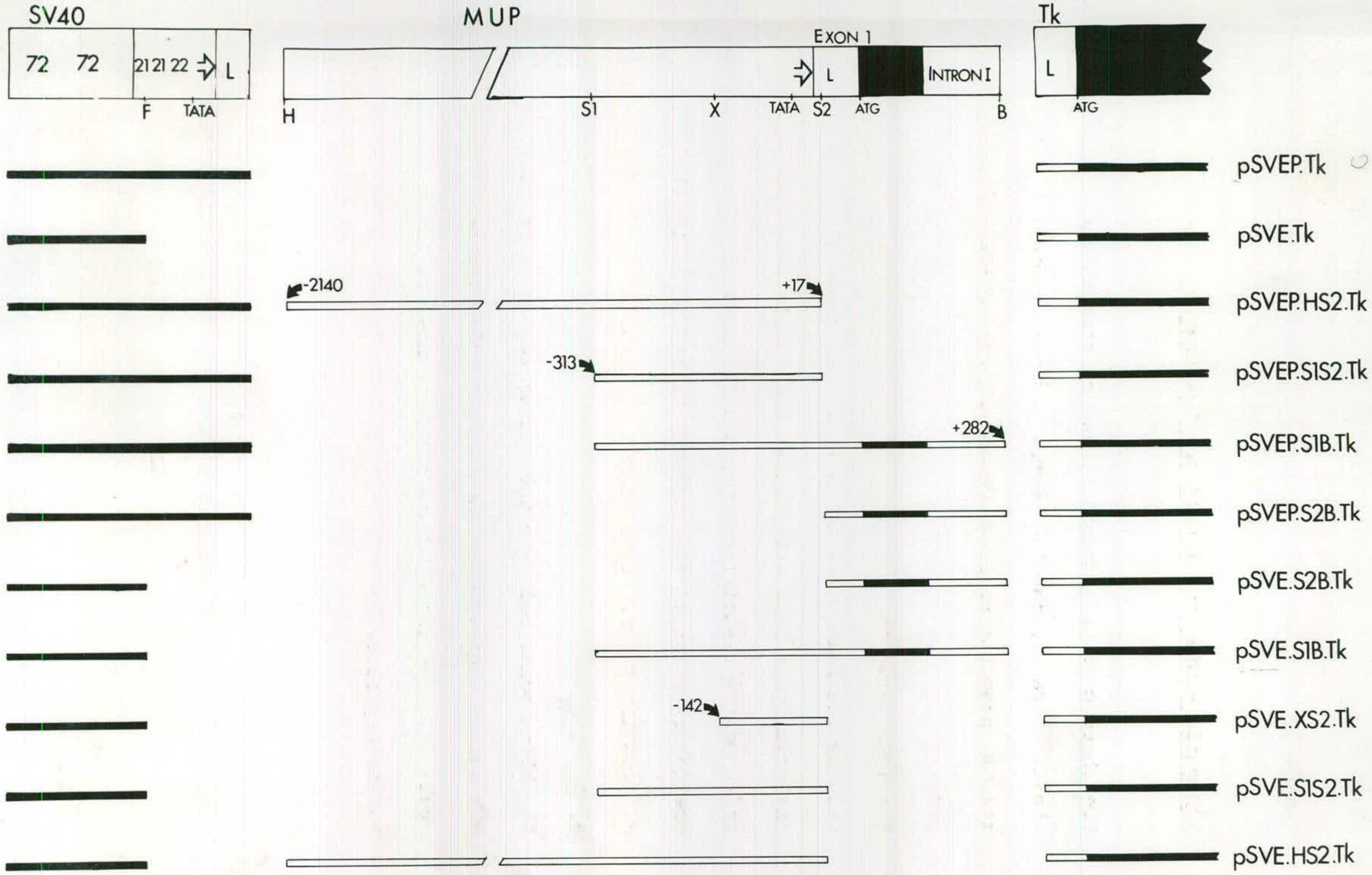
FIGURE 13

Schematic Representation of the Constructs

The three major components of the constructs are shown. These are the SV40 early promoter region, the MUP promoter region and part of the HSV-Tk gene.

The TATA boxes and ATG codons are indicated as well as the leader sequences (L). The arrows show the transcriptional start sites and the letters refer to the restriction enzyme sites used in the construction of the vectors. These are F, *Fok*I; H, *Hin*dIII, S1, S2, *Sau*3A sites; X, *Xmn*I and B, *Bam*HI.

The black boxes refer to the coding sequences in the MUP and HSV-Tk regions. The numbers refer to the nucleotide positions of the MUP sequences relative to its cap site.


SV40
72
72
21 21 22
L
F
TATA
MUP
EXON 1
L
INTRON I
H
S1
X
TATA
S2
ATG
B
Tk
L
ATG
-2140
+17
-313
+282
-142
pSVEP.Tk
pSVE.Tk
pSVEP.HS2.Tk
pSVEP.S1S2.Tk
pSVEP.S1B.Tk
pSVEP.S2B.Tk
pSVE.S2B.Tk
pSVE.S1B.Tk
pSVE.XS2.Tk
pSVE.S1S2.Tk
pSVE.HS2.Tk

sequencing, the RF was prepared and the fragment was gel purified from the vector after digestion with *Sma*I and *Hind*III. Subsequently, the appropriate digestion of the pSVEP. vectors allowed the replacement of the EP sequences with the E fragment, thus generating the pSVE. series. The pSVEP. and pSVE. vectors are referred to as those constructs containing the SV40 early promoter region and the SV40 enhancer, respectively. Briefly, pSVE.XS2.Tk was simply prepared by deleting the sequences between the *Xmn*I site and the *Sau*3A (S1 site) of the S1S2 fragment of BS6 in pSVEP.S1S2.Tk. The nomenclature adopted for the thymidine kinase gene and protein enzyme is Tk and tk respectively.

This chapter would not have been completed in time without the help given by Melville Richardson and Ann Duncan. My thanks to Melville for maintaining the stocks of plasmids and the plasmid preparations of pSVE.HS2.Tk and pSVE.XS2.Tk. Also many thanks to Ann for the preparation of RNA and their analysis on Northern blots.

FIGURE 14 and 15

Sequence structure of 14) the 5' end region of the Group 1 gene BS6 and 15) The SV40 early promoter region

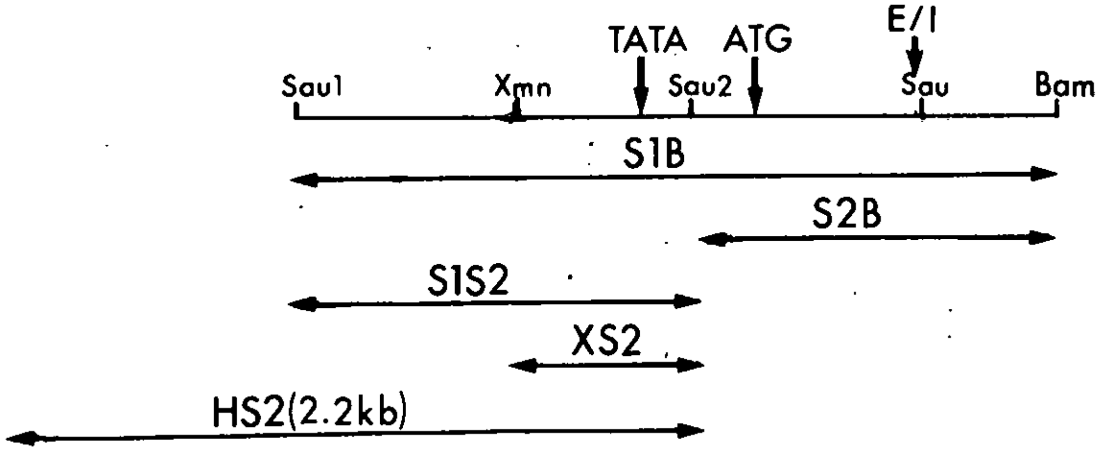
Underneath the sequences a schematic representation of the cloned regions used in the vector constructions is shown.

14) Arrows indicate those sequence symmetries found in BS6. The TATA box is underlined and the various restriction enzyme sites used in the construction of the vectors are shown.

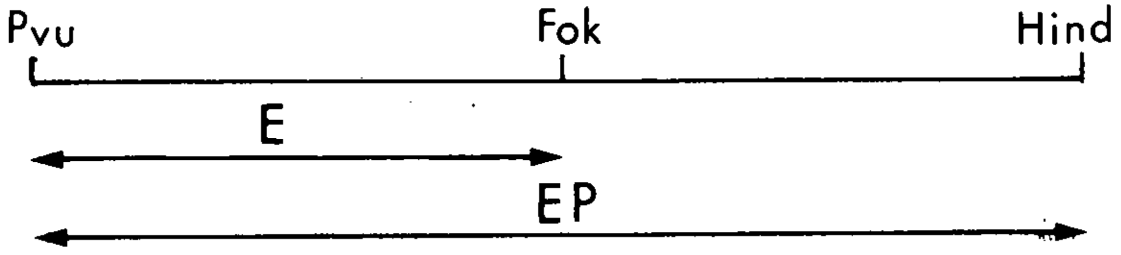
15) The dashed lines indicate the symmetries within the SV40 early promoter region while the vertical lines show the early early transcription start sites. The restriction enzyme sites used in the cloning of this region are shown above the sequences.

These diagrams were prepared by Dr. J. Bishop.

```
                                                    AGCTGGGT TCTCAGTAAA

 -550    AACAAGTTTC AAAGAATCTG ACTCCTTCAT CTGACCTCCG CTGCACCAGG

 -500    CCTGGACATG GTACTCAGAT ACACATGAAA AACCAGTGCT CATACACATG
                            |-------------->          |--------
 -450    AAATATGAAT GAATCTTTTG CAAAAATTGG GTAACTCAAG TTCTTCCATA
         ---->          |--------------->
 -400    CTCATCTTAA AGCAAAGATT CTATACATAC TTGTCCTTCC AGTTCAGTCT
             |--------------->                   Sau3A
 -350    GTACTCCCTC CAAAACTGGC TTTCTGATTC CAAACCAGAT CCAAAGTTGC
                                            <-----------|  |---
 -300    ATCTGGTTCC AGTTGGCAGT TCTCTTGAAC ACCCACTGTT TTATTGGGAA
         ------->
 -250    TATGTTTTGA GTAACATAAG ATTACTAAAT CAATCCATAG GTTATGAAAT

 -200    TGCCAAGTTT GCAAAGGGCA AGGAACAATT CTTGGCCTCT AATCAATAAA
                 XmnI
 -150    TGAGAAAACA TTCCACAAAG CCTGACAGAG GTAGAGGAGA CCCATACGGG

 -100    AAGAGGGAAA AAAAAAAAAA CAAAACAAAC AACAACAACA AAAAAAAAAA

  -50    ACCCGCTGAA CCCAGAGAGT ATATAAGGAC AAGCAAAGGG GCTGGGGAGT
                       Sau3A
    1    GGAGTGTAGC CACGATCACA AGAAAGACGT GGTCCTGACA GACAGACAAT

   51    CCTATTCCCT ACCAAAATGA AGATGCTGCT GCTGCTGTGT TTGGGACTGA  Exon 1

  101    CCCTAGTCTG TGTCCATGCA GAAGAAGCTA GTTCTACGGG AAGGAACTTT
                          Sau3A
  151    AATGTAGAAA AGGTATGATC ACTGAATAGT AGCTTCTGAC TCAGAATGTG

  201    CTTTGGGGAA CTCTTGAAGC CAAGTAGGTC CTTTGAGGGG ATGGGTATAG  Intron 1
                                    BamH1
  251    TGCCCCAATC TCTTAGACAA ATGAATGGAT CC
```



```
     1/2 PvuII
  1  CTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGC
                           ¦---------------------------

 51  AGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTG
     --------------------------------------------¦¦----

101  GAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTC
     --------------------------------------------------
                                   }--------FokI
151  AATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCT
     ---------------¦   ¦-------------------¦¦---------

201  AACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTT
     ----------¦ ¦-------------------¦
                            ¦¦      ¦¦¦>
251  TTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAGCTATTCCAGAAG
                      <------   ------>
                                            HindIII
301  TAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTT
```



ANALYSIS OF CONSTRUCTS.

The transcriptional activity of the various constructs described above were examined with thymidine kinase transient expression analysis after transfection into baby hamster kidney (BHKtk-) cells. These are a fibroblast cell line which have a deficient Tk gene and are therefore tk- (see Methods). After transfection into the BHKtk- cells and allowing a 20 hour adsorption period the cells were harvested, extracts was prepared and tk enzyme activity was determined. The time course of TMP synthesis was linear for 90 to 120 minutes. Each extract was sampled after 30, 60 and 120 minutes of incubation and the one hour point was taken. The number of counts recorded for the mock transformation was subtracted and the tk activity was expressed as the number of picomoles of dTMP synthesised per minute per milligram of soluble protein at 37^{o}C. In each experiment enzyme activity was normalised to the activity of an extract of cells transfected with pSVEP.Tk included as a positive control.

An example is shown in the Methods section and a summary of the relative tk activities of cells transfected with the pSVEP. and the pSVE. series of constructs is shown in Table 6. Northern blot and nuclease S1 protection analysis of transcripts from selected constructs transiently expressed in BHKtk- cells was performed.

Translational Effects

Reporter functions represent an indirect assay of the level of post-transcriptional activity. As well as transcriptional effects, post-transcriptional, translational and other effects determine the overall outcome of the experiment. Such effects must be taken into account when comparing different transfected recombinants which generate transcripts with slightly different sequences. Even though in this case most of the transcripts generated from different transfected DNAs are the same, other effects cannot be ruled out. Consequently, enzyme assays should only be used as indicators of potential effects.

In this study a few of the transfected recombinants generate mRNAs with different 5' sequences. Therefore, differences in their tk activities may be interpreted in terms of their different processing and translational efficiencies. These are the constructs with the S2B fragment of BS6 which contains the MUP first exon and about 100-bp of the first intron fused to the leader sequence of the HSV- Tk gene. A ten-fold reduction in tk activity is observed when the tk activity of cells transfected with these DNAs is compared with constructs which are identical except for the absence of the S2B MUP fragment (compare pSVEP.Tk and pSVEP.S2B.Tk, and also pSVEP.S1S2.Tk and pSVEP.S1B.Tk in Table 6). Nuclease S1 protection analysis of the transcripts generated from pSVEP.S2B.Tk and pSVEP.S1B.Tk show that they are not processed and therefore contain both the exonic and intronic sequences of the MUP gene (Figure 17). The MUP coding sequences read out of frame with the Tk coding

FIGURE 16.

Translation products of the 5' end of pSVEP.S2B (or S1B).Tk Constructs.

Lane c, is the MUP reading frame.

Lane a, is the HSV-Tk reading frame.

*, denotes the stop codons.

The underlined translation products show the open reading lengths.

```
        ¦--MUP-SIGNAL PEPTIDE---------->                            ¦--MUP
        ATGAAGATGCTGCTGCTGCTGTGTTTGGGACTGACCCTAGTCTGTGTCCATGCAGAAGAA

a        *  R  C  C  C  C  C  V  W  D  *  P  *  S  V  S  M  Q  K  K
b         E  D  A  A  A  A  V  F  G  T  D  P  S  L  C  P  C  R  R  S
c       M  K  M  L  L  L  L  C  L  G  L  T  L  V  C  V  H  A  E  E

        MATURE PROTEIN--------->                ¦-----INTRON I-------->
        GCTAGTTCTACGGGAAGGAACTTTAATGTAGAAAAGGTATGATCACTGAAATAGTAGCTT

a        L  V  L  R  E  G  T  L  M  *  K  R  Y  D  H  *  N  S  S  F
b         *  F  Y  G  K  E  L  *  C  R  K  G  M  I  T  E  I  V  A  S
c       A  S  S  T  G  R  N  F  N  V  E  K  V  *  S  L  K  *  *  L


        CTGACTCAGAATGTGTGCTTTGGGGAACTCTTGAAGCCAAGTAGGTCCTTTGAGGGGATG

a        *  L  R  M  C  A  L  G  N  S  *  S  Q  V  G  P  L  R  G  W
b         D  S  E  C  V  L  W  G  T  L  E  A  K  *  V  L  *  G  D  G
c       L  T  Q  N  V  C  F  G  E  L  L  K  P  S  R  S  F  E  G  M


                                            ¦-HSV Tk LEADER SEQUENCE
        GGTATAGTGCCCCAATCTCTTAGACAAATGAATGGATCTTGGTGGCGTGAAACTCCCGCA

a        V  *  C  P  N  L  L  D  K  *  M  D  L  G  G  V  K  L  P  H
b         Y  S  A  P  I  S  *  T  N  E  W  I  L  V  A  *  N  S  R  T
c       G  I  V  P  Q  S  L  R  Q  M  N  G  S  W  W  R  E  T  P  A


        ---------->                            ¦-tk MATURE PROTEIN--------->
        CCTCTTTGGCAAGCGCCTTGTAGAAGCGCGTATGGCTTCGTACCCCTGCCATCAACACGC

a        L  F  G  K  R  L  V  E  A  R  M  A  S  Y  P  C  H  Q  H  A
b         S  L  A  S  A  L  *  K  R  V  W  L  R  T  P  A  I  N  T  R
c       P  L  W  Q  A  P  C  R  S  A  Y  G  F  V  P  L  P  S  T  R


        GTCTGCGTTCGACCAGGCTGCGCGTTCTCGCGGCCATAGCAACCGACGTACGGCGTTGCG

a        S  A  F  D  Q  A  A  R  S  R  G  H  S  N  R  R  T  A  L  R
b         L  R  S  T  R  L  R  V  L  A  A  I  A  T  D  V  R  R  C  A
c       V  C  V  R  P  G  C  A  F  S  R  P  *  Q  P  T  Y  G  V  A
```

sequence and introduce numerous stop codons within the MUP intronic sequences (Figure 16). Although eukaryotic ribosomes initiate translation predominantly from the first AUG internal AUGs maybe utilized if, irrespective of the frame of the upstream AUG, termination signals occur between the two starts (Kozak, 1984b; Liu _et al_, 1984; Hunt, 1985). This in some situations may results in a reduced translational efficiency of the message (Kozak, 1984c; Liu _et al_, 1984). The next start codon which reads in frame with the Tk coding sequences occurs 112-bp downstream from the first stop codon. Use of this start codon would result in a 20 amino acid extension of the amino-terminal end of tk. Such tk might have reduced enzymatic activity. These considerations suggest that the reduced tk activity of pSVEP.S2B.Tk compared with pSVEP.Tk, and also pSVE(P).S1B.Tk compared with pSVE(P).S1S2.Tk may be due to the different translational efficiencies of their different mRNAs and/or to decreased activity of tk enzyme with a N-terminal extension.

Those contructs which generate identical transcripts with respect to the 5' noncoding, coding, and 3' noncoding sequences are presumed to be processed and translated identically. Differences in the tk activity of extracts of cells transfected with these various constructs are assumed to reflect differences in transcription.

PROMOTER UTILIZATION IN BHKtk- CELLS

pSVEP. Constructs

The SV40 enhancer is known to stimulate proximal promoters in preference to distal promoters (Wasylyk et al, 1983). Therefore, in cells transfected with the pSVEP. constructs transcripts originating from the SV40 promoter would be expected to pre-empt and occlude transcription from the MUP TATA box. Surprisingly, Northern blot analysis (Figure 18 and J.O. Bishop unpublished) indicate that the transcription observed is exclusively from the MUP promoter. S1 mapping of the 5' ends of transcripts synthesised from the pSVEP.Tk, pSVEP.S2B.Tk and pSVEP.S1B.Tk constructs transiently expressed in BHK cells, show that the SV40 early early promoter is used in pSVEP.Tk and pSVEP.S2B.Tk while pSVEP.S1B.Tk uses only the MUP promoter (Figure 17). These data show that the BS6 MUP TATA box is functional and that it is the preferred promoter of transcription even though the SV early promoter lies between it and the enhancer. Since the only transcripts observed are initiated from the MUP promoter it may be concluded that these sequences 'turn off' the SV40 promoter. The possibility that some transcription is initiated from the SV40 promoter under these circumstances cannot be ruled out. A formal confirmation might be obtained by means of run-on transcription experiments which would show the polymerase loading at the two promoter regions. This is however technically a very difficult experiment to perform (Treisman and Maniatis, 1985; Weber and Schaffner, 1985).

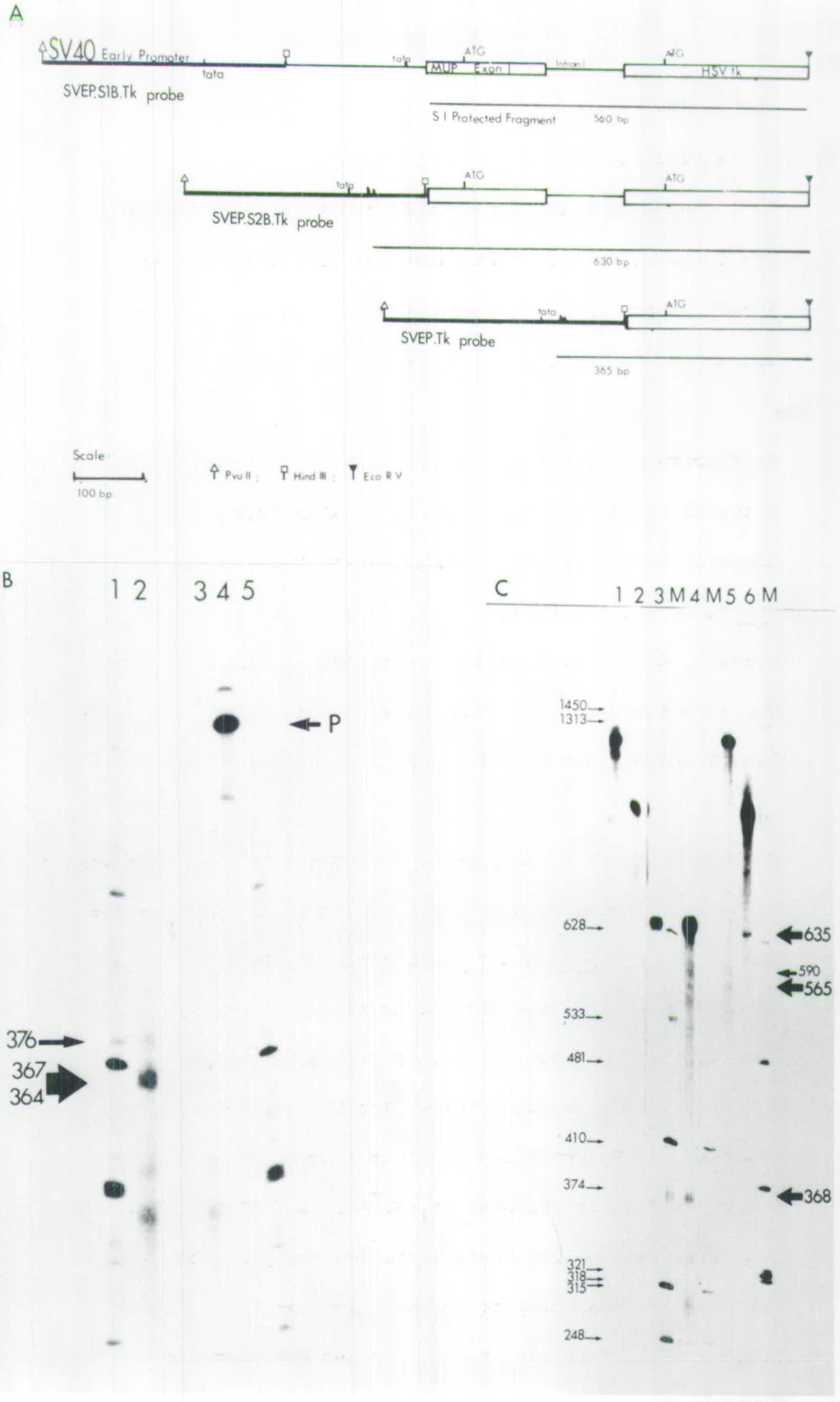
FIGURE 17.

S1 ANALYSIS OF mRNA SYNTHESISED IN BHK CELLS

A) A schematic representation of the S1 probes is shown. The lines below the probes show the protected fragment length of the probe. The probes were cloned into M13tg130 and uniformly labelled as described in Methods.

B) Electrophoretic analysis of the S1 mapping of the start sites of the SV40 early promoter of mRNA from cells transfected with pSVEP.Tk. Lanes 1 and 5 are markers (Sau3A digest of pBR322). 2, RNA from pSVEP.Tk transfected cells. 3, control - no RNA; 4, probe only (P). The arrows refer to the size of the protected fragments in lane 2 and map the SV40 early early initaition sites.

C) Electrophoretic analysis of S1 mapping of mRNA from BHK cells transfected with various constructs. M, denotes the marker tracks (TaqI and Sau3A digests of pBR322); lanes 1, 2, 3, are the control tracks for 5, 6, 4, - no RNA; 4, RNA from pSVEP.Tk; 5, pSVEP.S1B.Tk; and 6, pSVEP.S2B.Tk transfected cells. The 368 and 635 arrows indicate the S1 protected fragment lengths corresponding to the SV40 early promoter in pSVEP.Tk and pSVEP.S2B.Tk respectively. The 565 arrow indicates the protected fragment corresponding to the MUP promoter in pSVEP.S1B.Tk. This also shows that the MUP intronic sequences are not processed in pSVEP.S1B/S2B.Tk transfected cells.


A
SV40 Early Promoter
tata
tata
ATG
MUP Exon I
Intron I
ATG
HSV tk
SVEP.S1B.Tk probe
S I Protected Fragment
560 bp
tata
ATG
ATG
SVEP.S2B.Tk probe
630 bp
tata
ATG
SVEP.Tk probe
365 bp
Scale
100 bp
Pvu II ; Hind III ; Eco R V
B
1 2
3 4 5
P
376
367
364
C
1 2 3 M 4 M 5 6 M
1450
1313
628
533
481
410
374
321
318
315
248
635
590
565
368

Interestingly, at position -156 to -151 of BS6 a very good consensus of the polyadenylation signal (AATAAA) is present. This might argue that transcripts originating upstream of the MUP TATA box are terminated by processing. However, not all AATAAA sequences are functional and there is increasing evidence which suggests that the AATAAA motif is not the only signal required for the cleavage/polyadenylation event (Gil and Proudfoot, 1984). Another sequence found to be important is a G/T cluster situated about 30-bp downstream of the AATAAA signal (Birnstiel et al, 1985). Such a G/T cluster is not found in the vicinity of the -156 to -151 AATAAA motif of BS6. This would suggest that the -156 to -151 AATAAA motif is non-functional. Northern blot analysis of pSVEP.HS2.Tk transcripts probed with the whole plasmid show only those transcripts expected from the MUP promoter (Figure 18). If transcripts were originating at the SV40 promoter and terminating at -156 then a 2-kb transcript specific to the HS2 fragment would be expected. In fact only a 1.3-kb transcript is observed as is expected from transcription initiated at the MUP promoter.

pSVE. Constructs

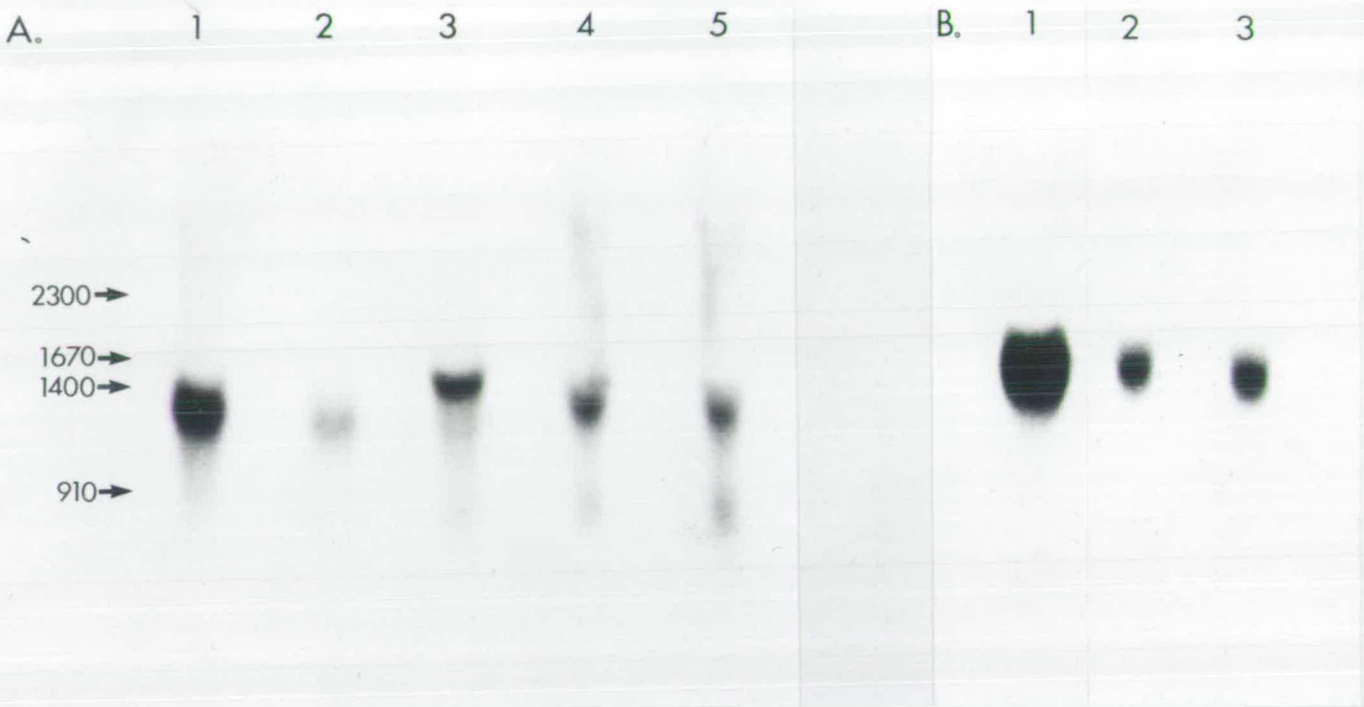
To investigate further whether the SV40 promoter has any significant effect on the MUP promoter activity, the pSVE. series of constructs was prepared. These constructs lack the SV40 promoter region and therefore should not show any influence this has on the MUP promoter. Although the above observations on the pSVEP. series suggest that the SV40 promoter is not used, transcription may still

FIGURE 18.

NORTHERN ANALYSIS OF mRNA SYNTHESISED IN BHK CELLS

A and B, are Northern blots of total mRNA synthesised in BHKtk- cells probed with Tk fragment (PstI - EcoRI) and pSVEP.HS2.Tk respectively. Markers are transferrin (2300), contrapsin (1670), alpha-1 antitrypsin (1400) and MUP (910).
The lanes refer to RNA from cells transformed with
A: 1, pSVEP.Tk; 2, pSVE.Tk; 3, pSVE.S2B.Tk; 4, pSVE.S1S2.Tk; and 5, pSVE.XS2.Tk. B: 1, pSVEP.Tk; 2, pSVEP.HS2.Tk and 3, pSVE.HS2.Tk.

C, shows a table of the predicted sizes of transcripts from the various regions of the constructs and the estimated sizes. This shows that pSVEP.HS2.Tk, pSVE.HS2.Tk, pSVE.S1S2.Tk and pSVE.XS2.Tk appear to initiate transcription from the MUP promoter region. While pSVE.Tk and pSVE.S2B.Tk appear to initiate transcription close to the SV40 enhancer region.



C.

SIZES OF mRNA SYNTHESISED IN BHKtk- DURING TRANSIENT EXPRESSION

CONSTRUCT	Predicted Size		Estimated Size
	SV-Region	MUP Promoter	
pSVEP.Tk	1350		1320
pSVE.Tk	1300		1300
pSVE.S2B.Tk	1550		1570
pSVEP.HS2.Tk	3650	1350	1320
pSVE.HS2.Tk	3600	1350	1320
pSVE.S1S2.Tk	1600	1350	1320
pSVE.XS2.Tk	1430	1350	1320

occur upstream even when the promoter is removed. Several studies have shown that the SV40 enhancer in the absence of its promoter can initiate transcription at random points just downstream from it and retain the same efficiency of transcription (Benoist and Chambon, 1981; Wasylyk et al, 1983). In this study those constructs lacking a promoter but retaining the enhancer (pSVE.Tk and pSVE.S2B.Tk) appear not to be an exception to this. Northern blot analysis indicates that transcription is initiated close to the enhancer element (Figure 18). Consequently, in cells transfected with the pSVE. constructs (containing the MUP promoter region) transcripts originating from sequences in the immediate vicinity of the enhancer would be expected to pre-empt and occlude transcription from the MUP promoter. Therefore, although a comparison between the pSVEP. and pSVE. series may indicate possible effects between the two promoters the pSVE. series is potentially just as complicated as the pSVEP. series. However, Northern blot analysis of pSVE.HS2.Tk, pSVE.XS2.Tk and pSVE.S1S2.Tk indicate that transcription occurs exclusively from the MUP promoter (Figure 18). This further supports the suggestion that the MUP promoter is functional as a consequence of its activation by the SV40 enhancer.

TABLE 6.

SUMMMARY OF tk ASSAYS DURING TRANSIENT EXPRESSION IN BHK

BHK tk$^-$ cells plated at 5 x 10^5 per 9cm dish and transfected with 10 micrograms of Form' I DNA.

	Relative tk activity(over mock).										
EXPERIMENT	1	2	3	4	5	6	7	8	9	10	Mean
CONSTRUCT											
pSVEP.Tk	100	100	100	100	100	100	100	100	100	100	100
pSVE.XS2.Tk									15 15	9 9	12.0
pSVE.S1S2.Tk		16 14	10 11	10 9		15 16		8 /	17 13	8 13	12.3
pSVE.HS2.Tk		10 10				16 13					12.3
pSVEP.S1S2.Tk		5 6		4 6		13 14		7 /			7.9
pSVEP.HS2.Tk		11 11	8 /			15 15					12.0
pSVE.Tk					30 34		31 30	28 /			30.6
pSVE.S1B.Tk		1 1			1 1		2 2				1.3
pSVE.S2B.Tk	14 8	4 6			11 11		18 12	15 /			11.0
pSVEP.S1B.Tk		0.1 0.5			0.5 0.5		1.5 1.5				0.8
pSVEP.S2B.Tk	15 10	5 6			14 /		14 16	9 10			11.0

T-TEST

COMPARISON	t value	dF	p
pSVE/P.S1S2.	3.333	18	0.005
pSVE/P.S1B.	2.387	10	0.05
pSVEP.HS2/S1S2.	2.890	10	0.025

COMPARISON OF tk ACTIVITIES

Low Efficiency of the MUP promoter.

The most striking observation to be drawn from Table 6 is the low efficiency of the MUP promoter under the effect of the SV40 control sequences. This efficiency compared to the SV40 early promoter (pSVEP.Tk) is one order of magnitude down and is specific to the MUP promoter sequences (compare pSVE(P).S2B.Tk with pSVE(P).S1B.Tk in Table 6).

Removal of the SV40 promoter.

In most cases the tk activity of the same MUP-reporter fragment is the same when associated with pSVEP. and pSVE., further supporting the suggestion that the SV40 promoter is not used. A slight drop in activity (a factor of 1.6) is observed when the SV40 promoter is linked to the S1 site of BS6 (compare pSVE.S1B.Tk and pSVEP.S1B.Tk, and also pSVE.S1S2.Tk and pSVEP.S1S2.Tk in Table 6). The fact that pSVEP.HS2.Tk and pSVE.HS2.Tk have equivalent levels of activity to pSVE.S1S2.Tk suggests that the drop in activity may be due to interference between the two promoters in such close proximity.

These data clearly show that the MUP promoter is active even when in the presence of an additional promoter with a more favourable position relative to the enhancer. The tentative suggestion that the MUP promoter 'switches off' the SV40 promoter and the fact that the MUP promoter in the absence of the enhancer is not functional, while in its presence it is used at a low efficiency, may implicate negative elements associated with the MUP promoter. The importance of negative regulation for the control of eukaryotic gene expression was discussed above. It may be argued that the MUP promoter is regulated in these cells and that the enhancer acts in some manner as an anti-repressor. The above data is consistant with this idea. However, it could be simply that *cis*-acting sequences necessary for the transcription in fibroblasts are not present within the MUP promoter region.

Resection of the MUP promoter linked to the SV40 Enhancer.

The distance between the enhancer and promoter may influence the enhancer effect (Wasylyk *et al*, 1984). However, resection of the MUP promoter from - 2140 to -142 show no significant changes in the level of tk expression (compare pSVE.HS2.Tk, pSVE.S1S2.Tk and pSVE.XS2.Tk in Table 6). If a negative sequence (silencer) is present within the MUP promoter region it must lie within the -142 region, since the removal of such a sequence would result in relief of expression. In this connection the A-Tract region of BS6 may be such a silencer (see Chapter 4).

Preliminary observations involving indirect competition assays with the MUP promoter sequences as the competitor, cotransformed with constructs which include the SV40 enhancer linked to the MUP promoter, show a two fold increase in expression of HSV-Tk. Although this is only a weak response it may imply the presence of a trans-acting repressor (data not shown).

CONCLUSIONS

A transient assay system for the MUP gene promoter functions in tissue culture cells has been developed. The system involves the transfection of rodent fibroblast cells (in this study BHKtk- cells) with constructs containing a MUP promoter (defined as the TATA box and associated sequences) which is linked to the HSV-Tk gene as the reporter function. The transient expression of HSV-Tk is only detected when such constructs are linked to the SV40 enhancer region (defined as the 72-bp repeats) or with the SV40 early promoter region (defined as the enhancer with the 21-bp repeats and TATA box). In the latter case transcription is only observed from the downstream MUP promoter and not from the SV40 early promoter. Removal of the SV40 21-bp repeats and TATA box while retaining the enhancer, has little or no effect on the expression of HSV-tk. Thus it is suggested that the SV40 early promoter may be switched off in the presence of the MUP promoter sequences. Resection of the MUP promoter, linked to the enhancer, from -2140 to -142 has no effect on the level of expression. However, the efficiency of these constructs is 10-fold down compared to that of the SV40 early promoter directly linked to the HSV-Tk gene.

Together these observations may suggest some sort of down regulatory effect on the SV40 early promoter by the MUP promoter while the SV40 enhancer has a deregulatory effect on the MUP promoter.

METHODS

Enzymes

Enzymes were obtained from New England Biolabs, Bethesda Research Laboratories and Amersham. All enzymes were used according to the manufacturer's recommendations.

Preparation of mouse male total mRNA

Total cellular RNA was prepared using the methodology of Chirgwin et al (1979). Liver tissue (about 1 g) was homogenised in 4 M guanidium thiocyanate, 0.5% w/v sodium N-lauryl sarcosinate, 25 mM sodium citrate, 0.1% w/v Sigma antifoam A and 0.1 M beta-mercaptoethanol in a Sorvall omnimix. The resulting solution was then centrifuged at 8k rpm, 10°C for 10 minutes (Sorvall HB4). The resulting pellet was discarded and the supernatant was layered onto 1.2 ml cushions of 5.7 M CsCl, 25 mM sodium acetate (pH 5) and centrifuged at 36k rpm, 20°C for 12 hours (Beckman SW50). The pelleted RNA was recovered by first carefully removing most of the overlaying solution with a Gilson pipette, drawing off the remainder and redissolving the RNA in 7.5 M guanidine HCl, 25 mM sodium citrate (pH7), 5 mM DTT. The RNA was then ethanol precipitated by the addition of 0.0025 volumes of acetic acid and 0.5 volumes of ethanol.

Fractionation of poly(A)+ RNA

Poly(A)+ RNA was fractionated by two cycles of oligo-dT cellulose chromatography essentially as described by Aviv and Leder (1972). The loading buffer was 0.5 M NaCl, 20 mM Tris-HCl pH7, 1 mM EDTA, 0.1% w/v SDS. Elution was in the same buffer without NaCl. RNA was recovered by ethanol precipitation.

Preparation of Nuclease S1 and Primer Extension Probes.

Annealing Primer: 2.5 ug of single-stranded M13 clone A22 (Clark *et al*, 1985a; see appendix) was mixed with 5 ng of sequencing primer (17-mer) and heated at 70°C for 7 minutes in 4 mM Tris-HCl pH7.9, 4 mM $MgCl_2$ and 20 mM NaCl in a total volume of 10 µl. This was allowed to cool for 30 minutes to room temperature in a beaker.

Synthesis of complementary strand: The annealing mix was taken to a volume of 25 ul containing 50 mM TrisHCl pH7.9, 2 mM $MgCl_2$, 10 mM DTT, 0.1 mM dGTP, dATP, dTTP, 40 uM dCTP, 10-20 µCi of alpha-^{32}P dCTP and 5-10 U of the Klenow fragment of DNA polymerase (Boehringer-Mannheim). The synthesis was carried out at 30°C for 90 minutes. This was then taken to *Eco*RI conditions and digested with *Eco*RI. The S1 probe fragment was gel purified from a 1.5% agarose gel. This was phenol extracted and stored under ethanol.

Primer extension probe: The procedure for labelling the primer extension probe was essentially as described above except that the clone used was S25 (see Clark et al (1985a) in the Appendix) and after the first extraction the probe was digested with Sau96I and the primer fragment gel purified on a 5% acrylamide gel.

S1 Mapping and Primer Extension of mRNA.

This was essentially as described in Clark et al (1985a) in the Appendix.

Cloning

Ligation of inserts into recombinant vectors was carried out with T4 DNA ligase at 14°C with a 3 molar excess of insert to vector DNA.

DNA Sequencing

DNA fragments to be sequenced were cloned into M13mp8 and 9 (Messing and Vieira, 1982) and M13tg131 (Kieny et al, 1983). The dideoxynucleotide sequencing method of Sanger et al (1977) was used to sequence the single-stranded templates essentially as described by Coulson and Winter (1982) except that alpha-^{32}P dCTP was substituted for alpha-^{32}P dATP and the synthetic universal primer (17-mer) was purchased from Uniscience. The systematic sequencing approach of Hong (1982) was essentially as described.

Plasmid DNA Preparations

HB101 was transfected with recombinant plasmids and grown with the appropriate selecting antibiotic. Transfections and isolation of plasmid DNA were carried out as described by Bishop, 1979; Bishop and Davies, 1980, except that the plasmid was passed over a Sepharose 2B (Pharmacia) column, developed with 0.3 M NaCl, 10 mM Tris-HCl pH 7.5, as a further purification step.

TISSUE-CULTURE METHODS

Cells and Media

The cells used in the work described here were BHKtk- cells (Kidney, Syrian or Golden hamster, Mesocricetus auratus). The parent line of these cells was derived from the kidneys of five unsexed, one-day-old hamsters in March 1961 (Macpherson and Stoker, 1962). The cells were grown in Dulbeccos Modified Eagle's medium (DMEM) supplemented with 10% foetal calf serum (Flow Laboratories), 100 U/ml penicillin and 100 U/ml streptomycin (Gibco). Culture vessels were plastic flasks and petri dishes (NUNC). Cells were given fresh medium or passaged once or twice per week and were detached from the vessels by trypsinisation as described by Spandidos and Wilkie (1984).

Transfection of BHKtk- Cells

The calcium phosphate precipitation procedure based on the method of Graham and van der Eb (1973) was used for introducing the DNA into the cells for the transient expression assays.

The protocol described by Spandidos and Wilkie (1984) was used except for the following changes: 1) The cells were plated one day prior to transformation in 4.5 or 9 cm diameter petri dishes at a density of 2 x 10^5 or 8 x 10^5 cells in 5 or 10 ml of medium. 2) 5 or 10 µg of DNA was added per 4.5 or 9 cm plate respectively. 3) Carrier DNA was omitted from the transfection. 4) After 20 hours contact with the precipitate the cells were harvested for tk assays. Replica transfections were done for each of the transfected recombinants. It was found that the level of tk expression after a 20 hours adsorption period was the same as after a 48 hours period with a glycerol shock (data not shown).

Thymidine Kinase Enzyme Assays

24 hours after the transformation stage, cells were washed twice with PBS and harvested by scraping into PBS. No detectable effects on tk activity were found when cells were left for up to 2 hours in PBS prior to centrifugation (data not shown). The cells were pelleted by centrifugation (2 min), washed in PBS and recentrifuged. The pellet was resuspended in 100 µl of 50 mM Tris-HCl (pH 7.5), 5 mM 2- mercaptoethanol and 5 µM thymidine. The cells were disrupted by sonication in a water bath at 4°C for 90

minutes and the cell debris pelleted by centrifugation in an Eppendorf centrifuge for 10 minutes.

The enzyme assay as described by Spandidos and Wilkie (1984) is based on measuring the conversion of radiolabelled thymidine to thymidine phosphate. The reaction conditions were the same as described by Spandidos and Wilkie (1984) except that 50 µl of lysed cells and a final concentration of 50 µCi/ml of tritium labelled thymidine was used. At 30, 90 and 120 minutes 25 µl of the reaction was spotted directly onto whatman DE81 paper discs. These were washed in 3 changes of 10 mM Tris.HCl pH7.5 with shaking and dried under a vacuum at 80°C for 40 minutes. The radioactivity on the filters was determined by liquid scintillation counting (scintillant was 5 g/l PPO, 0.3 g/l dimethyl POPOP in toluene). An example of an assay performed (Experiment 4, see Table 6) is shown in Figure 19.

The soluble protein content of the cell extracts was estimated by the method of Bradford (1976) using the Bio-Rad protein assay system. This is a dye-binding assay based on the differential colour change of a dye in response to various concentrations of protein.

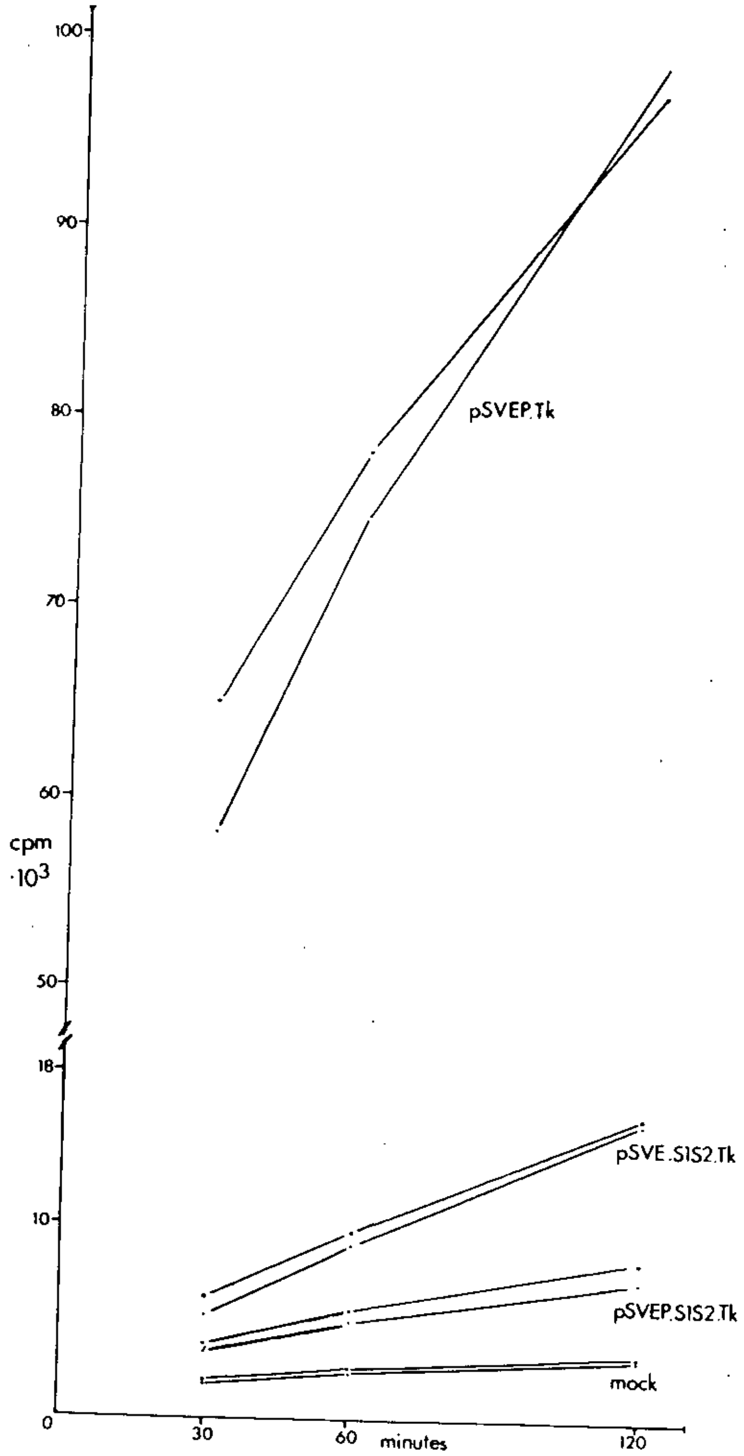
Nuclease S1 Protection Analysis of mRNA Extracted from Transfected Cells.

Transfections were performed as described above except that the amount of DNA loaded per transfection was 20 µg per 9 cm dish. Cells were harvested by scraping into 4 M guanidinium thiocyanate mix and total mRNA prepared as described above (see preparation of mouse

FIGURE 19.

Thymidine Kinase Activity from Cell Extracts Prepared from Transfected Cells. (Experiment 4).

A plot of the time course of the cpm of dTMP synthesised for the respective transfected cell extracts is shown. The Table below shows the amount of soluble protein per reaction sample, the 60 minute time point and the calculated enzyme activity. An average of the pSVEP.Tk extract enzyme activity was taken and all the other activities were related to this.



THYMIDINE KINASE ENZYME ACTIVITY

DNA Transfected	Soluble Protein Content of sample (µg)	60 min point less mock (cpm)	Enzyme activity pmol dTMP/min/mg soluble protein	Relative Efficiencies %
pSVEP.Tk	11	72620	14.3	100
pSVEP.Tk	12	76020	13.7	
pSVE.S1S2.Tk	11	6720	1.4	10
pSVE.S1S2.Tk	12	7320	1.3·	9
pSVEP.S1S2.Tk	9	3420	0.8	6
pSVEP.S1S2.Tk	8	2120	0.6	4

Reaction sample (25 µl) containing 7733 cpm/pmol thymidine.
Counting efficiency was 58%

male total mRNA).

Construction of Probes: pSVEP.Tk, pSVEP.S2B.Tk and pSVEP.S1B.Tk were digested with AvaI and PvuII and the 'probe' fragment isolated for each of the respective recombinants (see Figure 17). Each of these fragments was cloned into M13mp8 digested with SmaI and SstI. The cloning of the probe fragments were confirmed by sequencing.

Labelling of Probes: This was carried out essentially as described in Preparation of Nuclease S1 and Primer Extension Probes. After synthesis of the complementary strand the probes were digested with EcoRI (present in the polylinker) and EcoRV (present in the Tk gene). The S1 probe fragments were then gel purified from a 1.5% gel, phenol extracted and stored under ethanol.

Hybridization of Probes to mRNA: One sixth of the total cellular RNA extracted from the transfected cells was co-precipitated with 100,000 cmp of probe. The conditions were the same as described in the Appendix (Clark et al, 1985a) except that the temperature of hybridization was 52°C and the nuclease S1 was used at a concentration of 500 U/ml.

REFERENCES

Al-Shawi, R.A. (1985). PhD. Thesis, University of Edinburgh.

Aviv, H. and Leder, P. (1972). Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid cellulose. Proc. Natl. Acad. Sci. USA, 79: 1408-1412.

Banerji, J., Rusconi, S. and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27: 299-308.

Banerji, J., Olson, L. and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in the immunoglobulin heavy chain genes. Cell 33: 729-740

Barnes, W.M., Bevan, M. and Son, P.H. (1983). Kilo-sequencing: creation of an ordered nest of asymmetric deletions across a large target sequence carried on phage M13. in Wu, R., Grossman, L. and Moldave, K. (eds.) Academic Press. Methods in Enzymology 101: 98-122.

Bennett, K., Lalley, P., Barth, R. and Hastie, N. (1982). Mapping the structural genes coding for the major urinary proteins in the mouse: combined use of recombinant inbred strains and somatic cell hybrids. Proc. Natl. Acad. Sci. USA 79: 1220-1224.

Benoist, C. and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promoter region. Nature 290: 304- 310.

Bentley, D.L. and Rabbitts, T.H. (1983). Evolution of immunoglobulin V genes: Evidence indicating that recently duplicated human V-kappa sequences have diverged by gene conversion. Cell 32: 181-189.

Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983). Buffer gradient gels and 35-S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA 80:3963-3965.

Birnstiel, M.L., Busslinger, M. and Strub, K. (1985). Transcription termination and 3' processing: The end is in site! Cell 41: 349-359.

Bishop, J.O. (1979). A DNA sequence cleaved by restriction endonuclease R. EcoRI in only one strand. J. Mol. Biol. 125: 545-549.

Bishop, J.O. and Davies, J.A. (1980). Plasmid cloning vectors that can be nicked at a unique site. Molec. Gen. Genet. 179: 573-580.

Bishop, J.O., Clark, A.J., Clissold, P.M., Hainey, S., and Francke, U. (1982). Two main groups of mouse major urinary protein genes, both largely located on chromosome 4. EMBO. J. 1: 615-620.

Bishop, J.O., Selman, G.G., Hickman, J., Black, L., Saunders, R.D.P. and Clark, A.J. (1985). The 45-kb unit of major urinary protein gene organization is a gigantic imperfect palindrome. Mol. Cell. Biol. 5: 1591-1600.

Borgmeyer, U., Nowock, J. and Sippel, A.E. (1984). The TGGCA-binding protein: a eukaryotic nuclear protein recognizing a symmetrical sequence on double-stranded linear DNA. Nucl. Acid Res. 12: 4295-4311.

Borrelli, E., Hen, R. and Chambon, P. (1984). Adenovirus-2 E1A products repress enhancer-induced stimulation of transcription. Nature 312: 608-612.

Bradford, M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilising the principle of protein-dye binding. Anal. Biochem. 72: 248-254.

Brady, J., Radonovich, M., Vodkin, M., Natarajan, V., Thoren, M., Das, G., Janik, J. and Salzman, N.P. (1982). Site specific base substitution and deletion mutations that enhance or suppress transcription of the SV40 major late RNA. Cell 31: 625-633.

Breathnach, R. and Chambon, P. (1981). Organization and expression of eukaryotic split genes coding for protein. Ann. Rev. Biochem. 50: 349-383.

Bunick, D., Zandomeni, R., Ackeman, S. and Weinmann, R. (1982). Mechanisms of RNA polymerase II-specific initiation of transcription *in vitro*: ATP requirement and uncapped runoff transcripts. Cell *29*: 877-886.

Carter, A.D., Felber, B.K., Walling, M.J., Jubier, M.F., Schmidt, C.J. and Hamer, D.H. (1984). Duplicated heavy metal control sequences of the mouse metallothionein-I gene. Proc. Natl. Acad. Sci. USA. *81*: 7392-7396.

Cato, A.C.B., Geisse, S., Wenz, M., Westphal, H.M. and Beato, M. (1984). The nucleotide sequences recognized by the glucocorticoid receptor in the rabbit uteroglobin gene region are located far upstream from the initiation of transcription. EMBO J. *3*: 2771-2778.

Chamberlin, M.J. (1974). The selectivity of transcription. Ann. Rev. Biochem. *43*: 721-775.

Chandler, V.L., Maler, B.A. and Yamamoto, K.R. (1983). DNA sequences bound specifically by glucocorticoid receptor *in vitro* render a heterologous promoter hormone responsive *in vivo*. Cell *33*: 489-499.

Chirgwin, J.M., Przybyla, A.E., MacDonald, R.I., and Rutter, W.J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry *18*: 5294-5299.

Ciliberto, G., Dente, L. and Cortese, R. (1985). Cell-specific expression of a transfected human alpha 1-antitrypsin gene. Cell

41: 531-540.

Clark, A.J., Clissold, P.M. and Bishop, J.O. (1982). Variation between mouse major urinary protein genes isolated from a single inbred line. Gene 18: 221-230.

Clark, A.J., Clissold, P.M., Al-Shawi, R.A., Beattie, P. and Bishop, J.O. (1984a). Structure of mouse major urinary protein genes: different splicing configurations in the 3' non-coding region. EMBO. J. 3: 1045-1052.

Clark, A.J., Hickman, J. and Bishop, J.O. (1984b). A 45-kb DNA domain with two divergently orientated genes is the unit of organisation of the murine major urinary protein genes. EMBO. J. 3: 2055-2064.

Clark, A.J., Ghazal, P., Bingham, R.W., Barrett, D. and Bishop, J.O. (1985a). Sequence structure of a mouse major urinary protein gene and pseudogene compared. EMBO. J. 4: 3159-3165.

Clark, A.J., Chave-Cox, A., Ma, X. and Bishop, J.O. (1985b). Analysis of major urinary protein genes: variation between the exonic sequences of group 1 genes and a comparison with an active gene outwith group 1 both suggest that gene conversion has occurred between MUP genes. EMBO. J. 4: 3167-3171.

Clayton, D.E. and Darnell, Jr., J.E. (1983). Changes in liver-specific compared to common gene transcription during primary culture of mouse hepatocytes. Mol. Cell. Biol. 3: 1552-1561

Clissold, P.M. and Bishop, J.O. (1982). Variation in mouse major urinary protein (MUP) genes and the MUP gene products within and between inbred lines. Gene 18: 211-220.

Clissold, P.M., Hainey, S. and Bishop, J.O. (1984). Messanger RNAs coding for mouse major urinary proteins are differentially induced by testosterone. Biochem. Genet. 22: 379-387.

Close, T.J., Christmann, J.L. and Rodriguez, R.L. (1983). M13 bacteriophage and pUC plasmids containing DNA inserts but still capable of beta-galactosidase alpha-complementation. Gene 23: 131-136.

Coulson, A. and Winter, G. (1982). Chain terminator sequencing course manual. MRC Centre, Cambridge.

Derman, E. (1981). Isolation of a cDNA clone for mouse urinary proteins: Age- and sex-related expression of mouse urinary protein genes is transcriptionally controlled. Proc. Natl. Acad. Sci. USA 78: 5425-5429.

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programmes for the VAX. Nucl. Acid Res. 12: 387-395.

Dierks, P., van Voyen, A., Cochran, M.D., Dobkin, C., Reiser, J. and

Weissmann, C. (1983). Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit beta-globin gene in mouse 3T6 cells. Cell 32: 695-706.

Dolan, K.P., Unterman, R., McLaughlin, M., Nakhasi, H.L., Lynch, K.R. and Feigelson, P. (1982). The structure and expression of very closely related members of the alpha-2u globulin gene family. J. Biol. Chem. 257: 13527-13534.

Dudler, R. and Travers, A.A. (1984). Upstream elements necessary for optimal function of the hsp70 promoter in transformed flies. Cell 38: 391-398.

Dudov, K.P. and Perry, R.P. (1984). The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron containing gene and an unmutated unprocessed gene. Cell 37:475-468.

Durnam, D.M. and Palmiter, R.D. (1981). Transcriptional regulation of the mouse metallothionein -I gene by heavy metals. J. Biol. Chem. 256: 5712-5716.

Dynan, W.S. and Tjian, R. (1983). The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. Cell 35: 79-87.

Dynan, W.S., Saffer, J.D., Lee, W.S. and Tjian, R. (1985). Transcription factor Sp1 recognizes promoter sequences from the monkey genome that are similar to the simian virus 40 promoter. Proc. Natl. Acad. Sci. USA 82: 4915-4919.

Dynan, W.S., Sazer, S., Tjian, R. and Schimke, R.T. (1986). Transcription factor Sp1 recognizes a DNA sequence in the mouse dihydrofolate reductase promoter. Nature 319: 246-248.

Elgin, S.C.R. (1984). Anatomy of hypersensitive sites. Nature 309: 213-214.

Emerson, B.M. and Felsenfeld, G. (1984). Specific factor conferring nuclease hypersensitivity at the 5' end of the chicken adult beta-globin gene. Proc. Natl. Acad. Sci. USA. 81: 95-99.

Emerson, B.M.; Lewis, C.D. and Felsenfeld, G. (1985). Interaction of specific nuclear factors with the nuclease-hypersensitive region of the chicken adult beta-globin gene: Nature of the binding domain. Cell 41: 21-30.

Episkopou, V., Murphy, A.J.M. and Efstratiadis, A. (1984) Cell-specified expression of a selectable hybrid gene. Proc. Natl. Acad. Sci. USA. 81: 4657-4661.

Finlayson, J.S., Asofsky, R., Potter, M. and Runner, C.C. (1965). Major urinary protein complex of normal mice: origin. Science 149: 981-982.

Foster, J.; Stafford, J. and Queen, C. (1985). An immunoglobulin promoter displays cell-type specificity independently of the enhancer. Nature 315: 423-425.

Ghazal, P., Clark, A.J. and Bishop, J.O. (1985). Evolutionary amplification of a pseudogene. Proc. Natl. Acad. Sci. USA. 82: 4182-4185.

Ghosal, D. and Saedler, H. (1978). DNA sequence of the mini-insertion IS2-6 and its relation to the sequence of IS2. Nature 275: 611-617.

Ghosh, P.K., Lebowitz, P., Frisque, R.J. and Gluzman, Y. (1981). Identification of a promoter component involved in positioning the 5' termini of simian virus 40 early mRNAs. Proc. Natl. Acad. Sci. USA 78: 100-104.

Gidoni, D., Dynan, W.S. and Tjian, R. (1984). Multiple specific contacts between a mammalian transcription factor and its cognate promoters. Nature 312: 409-413.

Gidoni, D., Kadonaga, J.T., Barrera-Saldana, H., Takashashi, K., Chambon, P. and Tjian, R. (1985). Bidirectional SV40 transcription mediated by tandem Sp1 binding interactions. Science 230: 511-517.

Gil, A. and Proudfoot, N.J. (1984). A sequence downstream of AAUAAA is required for rabbit beta-globin mRNA 3'-end formation. Nature 312: 473-474.

Gillies, S.D., Morrison, S.L., Oi, V.T. and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell 33: 717-728.

Goodboum, S., Zinn, K. and Maniatis, T. (1985). Human beta-interferon gene expression is regulated by an inducible enhancer element. Cell 41: 509-520.

Gorman, C.M., Rigby, P.W.J. and Lane, D.P. (1985). Negative regulation of viral enhancers in undifferentiated embryonic stem cells. Cell 42: 519-526.

Graham, F.L. and van der Eb, A.J. (1973). A new technique for the assay of infectivity of human adenovirus 5 DNA. Virology 52: 456-467.

Grosveld, G.C., de Boer, E., Shewmaker, C.K. and Flavell, R.A. (1982). DNA sequences necessary for transcription of the rabbit beta-globin gene *in vivo*. Nature 295: 120-126.

Gruss, P., Dhar, R. and Khoury, G. (1981). Simian virus 40 tandem repeated sequences as an element of the early promoter. Proc. Natl. Acad. Sci. USA 78: 943-947.

Gubits, R.M., Lynch, K.R., Kulkarni, A.B., Dolan, K.P., Gresik, E.W. and Feigelson, P. (1984). Differential regulation of alpha-2u globulin gene expression in liver, lachrymal gland, and salivary gland. J. Biol. Chem. 259: 12803-12809.

Hastie, N.D. and Held, W.A. (1978). Analysis of mRNA populations by cDNA.mRNA hybrid-mediated inhibition of cell-free protein synthesis. Proc. Natl. Acad. Sci. USA 75: 1217-1221.

Hastie, N.D., Held, W.A. and Toole, J.J. (1979). Multiple genes coding for the androgen-regulated major urinary proteins of the mouse. Cell 17: 449-457.

Henikoff, S. (1985). One-way exonucleolytic digestion to target deletion breakpoints for DNA sequencing. Gene 28: 351-359.

Hennighausen, L., Siebenlist, U., Danner, D., Leder, P., Rawlins, D., Rosenfeld, P. and Kelly, T.Jr. (1985). High-affinity binding site for a specific nuclear protein in the human IgM gene. Nature 314: 289-292.

Hentschel, C.C. and Birnstiel, M.L. (1981). The organization and expression of histone gene families. Cell 25: 301-313.

Holmgren, R., Corces, V., Morimoto, R., Blackman, R. and Meselson, M. (1981). Sequence homologies in the 5' regions of four *Drosophila* heat-shock genes. Proc. Natl. Acad. Sci. USA. 78: 3775-3778.

Hong, G.F. (1982). A systematic DNA sequencing method. J. Mol. Biol. 158: 539-549.

Humphries, R.K., Ley, T., Turner, P., Moulton, A.D. and Nienhuis, A.W. (1982). Differences in human alpha-, beta- and delta-globin gene expression in monkey kidney cells. Cell *30*: 173-183.

Hunt, T. (1985). False starts in translational control of gene expression. Nature *316*: 580-581.

Iatrou, K. and Tsitilou, S.G. (1983). Coordinately expressed chorion genes in *Bombyx mori*: is developmental specificity determined by secondary structure recognition. EMBO J. *2*: 1431-1440.

Ishihara, T., Kudo, A. and Watunabe, T. (1984). Induction of immunoglobulin gene expression in mouse fibroblasts by cyclohexamide treatment. J. Exp. Med. *160*: 1937-1942.

Jackson, J.A. and Fink, G.R. (1981). Gene conversion between duplicated genetic elements in yeast. Nature *292*: 306-311.

Jones, C.W. and Kafatos, F.C. (1980). Structure, organization and evolution of developmentally regulated chorion genes in a silkmoth. Cell *22*: 855-867.

Jones, K.A. and Tjian, R. (1985). Sp1 binds to promoter sequences and activates HSV 'immediate-early' gene transcription *in vitro*. Nature *317*: 179-183.

Karin, M. and Richards, R.I. (1982). Human metallothionein genes- primary structure of the metallothionein II gene and a related processed gene. Nature 229: 797-802.

Karin, M., Haslinger, A., Holtgrev, H., Richards, R.I., Krauter, P., Westphal, H.M. and Beato, M. (1984). Characterization of DNA sequences through which cadmium and glucocorticoid hormones induce human metallothionein-IIA gene. Nature 308: 513-519.

Karlin, S., Ghandour, G., Ost, F., Tavare, S. and Korn, L.J. (1983). New approaches for computer analysis of nucleic acid sequences. Proc. Natl. Acad. Sci. USA. 80: 5660-5664.

Khoury, G. and Gruss, P. (1983). Enhancer elements. Cell 33: 313-314.

Kieny, M.P., Lathe, R. and Lecocq, J.P. (1983). New versatile cloning and sequencing vectors based on bacteriophage M13. Gene 26: 91-99.

Killary, A.M. and Fournier, R.E.K. (1984). A genetic analysis of extinction: Trans-dominant loci regulate expression of liver-specific traits in hepatoma hybrid cells. Cell 38: 523-534.

Klein, H.L. and Petes, T.D. (1981). Intrachromosomal gene conversion in yeast. Nature 289: 144-148.

Knopf, J.L., Gallagher, J.F. and Held, W.A. (1983). Differential, multihormonal regulation of the mouse major urinary protein gene family in the liver. Mol. Cell. Biol. 3: 2232-2240.

Konarska, M.M., Padgett, R.A. and Sharp, P.A. (1984). Recognition of cap structure in splicing in vitro of mRNA precursors. Cell 38: 731-736.

Kozak, M. (1984a). Compilation and analysis of sequences upstream from the translational start site in eukarotic mRNAs. Nucleic Acids Res. 12: 857-872.

Kozak, M. (1984b). Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. Nature 308: 241-246.

Kozak, M. (1984c). Selection of initiation sites by eukaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequences for preproinsulin. Nucleic Acids Res. 12: 3873-3893.

Krauter, K., Leinwand, L., D'Eustachio, P., Ruddle, F. and Darnell, J.E.Jr. (1982). Structural genes of the mouse major urinary protein are on chromosome 4. J. Cell. Biol. 94: 414-417.

Kuhn, N.J., Woodworth-Gutai, M., Gross, K.W. and Held, W.A. (1984). Subfamilies of the mouse major urinary protein (MUP) multi-gene family: sequence analysis of cDNA clones and differential regulation in the liver. Nucl. Acid Res. 12: 6073-6090.

Kulkarni, A.B., Gubits, R.M. and Feigelson, P. (1985). Developmental and hormonal regulation of alpha-2u globulin gene transcription. Proc. Natl. Acad. Sci. USA. 82: 2579-2582.

Kurtz, D.T. (1981a). Rat alpha-2u globulin is encoded by a multigene family. J. Mol. Appl. Genet. 1: 29-38.

Kurtz, D.T. (1981b). Hormonal inducibility of rat alpha-2u globulin genes in transfected mouse cells. Nature 291: 629-631.

Lacy, E. and Maniatis, T. (1980). The nucleotide sequence of a rabbit beta-globin pseudogene. Cell 21: 545-553

Laimins, L.A., Khoury, G., Gorman, C., Howard, B. and Gruss, P. (1982). Host specific activation of transcription by tandem repeats from simian virus 40 and Moloney murine sarcoma virus. Proc. Natl. Acad. Sci. USA. 79: 6453-6457.

Laperche, Y., Lynch, K.R., Dolan, K.P. and Feigelson, P. (1983). Tissue-specific control of alpha-2u globulin gene expression: constitutive synthesis in the submaxillary gland. Cell 32: 453-460.

Lee, D.C. and Roeder, R.G. (1981). Transcription of adenovirus type 2 genes in a cell-free system: Apparent heterogeneity of initiation at some promoters. Mol. Cell. Biol. 1: 635-651.

Lee, M.G-S., Lewis, S.A., Wilde, C.D., and Cowan, M.J. (1983).

Evolutionary history of a multigene family: An expressed beta-tubulin gene and three processed pseudogenes. Cell 33: 477-482.

Levinger, L. and Varshavsky, A. (1982). Protein D1 preferentially binds A + T rich DNA *in vitro* and is a component of *Drosophila melanogaster* nucleosomes containing A + T rich satellites DNA. Proc. Natl. Acad. Sci. USA. 79: 7152-7156.

Liu, C.-C., Simonsen, C.C. and Levinson, A.D. (1984). Initiation of translation at internal AUG codons in mammalian cells. Nature 309: 82-85.

Lomonossoff, G.P., Butler, P.J.G. and Klug, A. (1981). Sequence dependent variation in the conformation of DNA. J. Mol. Biol. 149: 745-760.

Lynch, K.R., Dolan, K.P., Nakhasi, H.L., Unterman, R. and Feigelson, (1982). The role of growth hormone in alpha-2u globulin synthesis: a reexamination. Cell 28: 185-189.

McIntyre, K.R. and Seidman, J.G. (1984). Nucleotide vsequence of mutant 1-A beta bm12 gene is evidence for genetic exchange between mouse immune response genes. Nature 308: 551-553,

McKnight, S.L. (1982). Functional relationship between transcriptional control signals of the thymidine kinase gene of herpes simplex virus. Cell 31: 355-365.

McKnight, S.L., Kingsbury, R.C., Spence, A. and Smith, S. (1984). The distal transcription signals of the herpesvirus tk gene share a common hexanucleotide control sequence. Cell 37: 253-262.

Macpherson, I. and Stoker, M. (1962). Polyoma trsansformation of hamster cell clones - an investigation of genetic factors affecting cell competence. Virology 16: 147-151.

Melon, P., Parker, V., Gluzman, Y. and Maniatis, T. (1981). Identification of DNA sequences required for transcription of the human alpha 1 -globin gene in a new SV40 host-vector system. Cell 27: 279-288.

Melton, D.W., McEwan, C., McKie, A.B. and Reid, A.M. (1986). Expression of the mouse HPRT gene: Deletional analysis of the promoter region of an X-chromosome linked housekeeping gene. Cell 44: in press.

Mercola, M., Wang, X-F., Olsen, J. and Calame, K. (1983). Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus. Science 221: 663-665.

Messing, J. and Vieira, J. (1982). A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene. 19: 269-276.

Motwani, N.M., Unakar, N.J. and Roy, A.K. (1980). Multiple hormone requirement for the synthesis of alpha-2u globulin by monolayers of rat hepatocytes in long term primary culture. Endocrinology 107: 1606-1613.

Myers, R.M., Rio, D.C., Robbins, A.K. and Tjian, R. (1981). SV40 gene expression is mediated by the cooperative binding of T-antigen to DNA. Cell 25: 373-384.

Nagata, K., Guggenheimer, R.A. and Hurwitz, J. (1983). Specific binding of a cellular DNA replication protein to the origin of replication of adenovirus DNA. Proc. Natl. Acad. Sci. USA. 80: 6177-6181.

Nabeshima, Y., Fujii-Kunayawa, Y., Muramatsu, M. and Ogata, K. (1984). Alternative transcription and two modes of splicing result in two myosin light chains from one gene. Nature 308: 333-338.

Neuberger, M.S. (1983). Expression and regulation of immunoglobulin heavy chain gene transfected into lymphoid cells. EMBO J 2: 1373-1378.

Norstedt, G., and Palmiter, R.D. (1984). Secretory rhythm of growth hormone regulates sexual differentiation of mouse liver. Cell 36: 805-812.

Nowock, J., Borgmeyer, U., Puschel, A., Rupp, R.A.W. and Sippel, A. E. (1985). The TGGCA protein binds to the MMTV-LTR, the adenovirus origin of replication, and the BK virus enhancer. Nucl. Acid Res. 13: 2045-2061.

Ollo, R. and Rougeon, F. (1983). Gene conversion and polymorphism: Generation of mouse immunoglobulin alpha 2a chain alleles by differential gene conversion by gamma 2b chaion gene. Cell 32: 515-523.

Osborne, T.F., Goldstein, J.L. and Brown, M.S. (1985). 5' end of HMG CoA reductase gene contains sequences responsible for cholesterol-mediated inhibition of transcription. Cell 42: 203-212.

Ott, M.O., Sperling, L., Herbomel, P., Yaniv, M. and Weiss, M.C. (1984). Tissue-specific expression is conferred by a sequence from the 5'end of the rat albumin gene. EMBO. J. 3: 2505-2510.

Parker, C.S. and Topol, J. (1984). A Drosophila RNA polymerase II transcription factor binds to the regulatory site of an hsp70 gene. Cell 37: 273-283

Pelham, H.R.B. (1982). A regulatory upstream promoter element in the Drosophila hsp70 heat-shock gene. Cell 30: 517-528.

Pelham, H.R.B., and Bienz, M. (1982). A synthetic heat- shock promoter element confers heat-inducibility of the herpes simplex thymidine kinase gene. EMBO J. 1: 1473-1477.

Pervaiz, S. and Brew, K. (1985). Homology of beta-lactoglobulin, serum retinol-binding protein, and protein HC. Science 228: 335-337.

Picard, D. and Schaffner, W. (1984). A lymphocyte-specific enhancer in the mouse immunoglobulin kappa gene. Nature 307: 80-82.

Poncz, M., Solowiejczyk, D., Ballantine, M., Schwartz, E. and Surrey, S. (1982). "Nonrandom" DNA sequence analysis in bacteriophage M13 by the dideoxy chain-termination method. Proc. Natl. Acad. Sci. USA. 79: 4298-4302.

Proudfoot, N.J. and Maniatis, T. (1980). The structure of a human alpha-globin pseudogene and its relationship to alpha-globin gene duplication. Cell 21: 537-545.

Queen, C. and Baltimore, D. (1983). Immunoglobulin gene transcription is activated by downstream sequence elements. Cell 33: 741-748.

Queen, C. and Stafford, J. (1984). Fine mapping of an Immunoglobulin gene activator. Mol. Cell. Biol. 4: 1042-1049.

Radding, C.M. (1978). Genetic recombination: strand transfer and mismatch repair. Ann. Rev. Biochem. 47: 847-880.

Ragg, H. and Weissmann, C. (1983). Not more than 117bp of 5' flanking sequence are required for inducible expression of a human IFN-alpha gene. Nature 303: 439-442.

Rawlins, D.R., Rosenfeld, P.J., Wides, R.T., Callberg, M.D. and Kelly, T.J.,Jr. (1984). Structure and function of the adenovirus origin of replication. Cell 37: 309-319.

Renkawitz, R., Beng, H., Graf, T., Matthias, P., Grez, M. and Schutz, G. (1982). Expression of a chicken lysozyme recombinant gene is regulated by progesterone and dexamethasone after microinjection into oviduct cells. Cell 31: 167-176.

Renkawitz, R., Schutz, G., van der Ahe, D. and Beato, M. (1984). Sequences in the promoter region of the chicken lysozyme gene required for steroid regulation and receptor binding. Cell 37: 503-510.

Reynolds, G.A., Basu, S.K., Osborne, T.F., Chin, D.J., Gil, G., Brown, M.S., Goldstein, J.L. and Luskey, K.L. (1984). HMG CoA reductase: a negatively regulated gene with unusual promoter and 5' untranslated regions. Cell 38: 275-285.

Roy, A.K., Demyan, W.F., Majumdar, D., Murty, C.V.R. and Chatterjee, B. (1983). Age-dependent changes in the androgen sensitivity of rat liver. In Steroid Hormone Receptors: Structure and Function, eds. Eriksson, H. and Gustafsson, J.A., 439-459. Elsevier Science Publishers B.V.

Ryals, J., Dierks, P., Ragg, H. and Weissmann, C. (1985). A 46-nucleotide promoter segment from an IFN-alpha gene renders an unrelated promoter inducible by virus. Cell 41: 497-507.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. 74: 5463-5467.

Saucier, J.M. and Wang, J.C. (1972). Angular alteration of the DNA helix by E. coli RNA polymerase. Nature New Biol. 239: 167-170.

Scheidereit, C., Geisse, S., Westphal, H.M. and Beato, M. (1983). The glucocorticoid receptor binds to defined nucleotide sequences near the promoter of mouse mammary tumour virus. Nature 304: 749-752.

Scherer, S. and Davis, R.W. (1980). Recombination of dispersed repeated DNA sequences in yeast. Science209: 1380-1384.

Scholer, H.R. and Gruss, P. (1984). Specific interaction between enhancer-containing molecules and cellular components. Cell 36:

403-411.

Searle, P.F., Davison, B.L., Stuart, G.W., Wilkie, T.M., Norstedt, G. and Palmiter, R.D. (1984). Regulatiuon, linkage and sequence of mouse metallothionein I and II genes. Mol. Cell. Biol. 4: 1221-1230.

Searle, R.F., Stuart, G.W. and Palmiter, R.D. (1985). Building a metal-responsive promoter with synthetic regulatory elements. Mol. Cell. Biol. 5: 1480-1489.

Seguin, C., Felber, B.K., Carter, A.D. and Hamer, D.H. (1984). Competition of cellular factors that activate metallothionein gene transcription. Nature 312: 781-785.

Shahan, K. and Derman, E. (1984). Tissue-specific expression of major urinary protein (MUP) genes in mice: Characterization of MUP RNAs by restriction mapping cDNA and by *in vitro* translation. Mol. Cell. Biol. 4: 2259-2265.

Shaw, P.H., Held, W.A. and Hastie, N.D. (1983). The gene family for major urinary proteins: expression in several secretory tissues of the mouse. Cell 32: 755-761.

Siebenlist, U., Hennighausen, L., Battey, J. and Leder, P. (1984). Chromatin structure and protein binding in the putative regulatory region of the c-*myc* gene in Burkitt Lymphoma. Cell 37: 381-391.

Slightom, J.L., Blechl, A.G. and Smithies, O. (1980). Human fetal G gamma- and A gamma- globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell21: 627-638.

Smith, G.P. (1976). Evolution of repeated DNA sequences by unequal crossover. Science 191: 528-535.

Solomon, M.J., Strauss, F. and Varshavsky, A. (1986). A mammalian HMG protein (alpha-protein) recognizes any stretch of six A.T base pairs in duplex DNA. Proc. Natl. Acad. Sci. USA. 86: in press.

Spandidos, D.A. and Wilkie, N.M. (1984). in Transcription and translation: A practical approach. Edited by Hames, B.D. and Higgins, S.J. IRL Press, Oxford. pp. 1-48.

Strauss, F. and Varshavsky, A. (1984). A protein binds to a satellite DNA repeat at three specific sites that would be brought into mutual proximity by DNA folding in the nucleosome. Cell 37: 889-901.

Stuart, G.W., Searle, R.F., Chen, H.Y., Brinster, R.L. and Palmiter, R.D. (1984). A 12 base pair DNA motif thast is repeated several times in metallothionein gene promoters confers metal regulation to a heterologous gene. Proc. Natl. Acad. Sci. USA. 81: 7318-7322.

Szoka, P. and Paigen, K. (1978). Regulation of mouse major urinary production by the Mup-a gene. Genetics 90: 597-612.

Tjian, R. (1981). T antigen binding and the control of SV40 gene expression. Cell 26: 1-2.

Tooze, J. (1980) DNA Tumor Viruses. Molecular Biology of Tumor Viruses, Part 2, revised. (CSHL).

Treisman, R. and Maniatis, T. (1985). Simian virus 40 enhancer increases number of RNA polymerase II molecules on linked DNA. Nature 315:72-75.

Vandenbergh, J.G., Finlayson, J.S., Dobrogosz, W.S., Dills, S.S. and Kost, T.A. (1976). Chromatographic separation of puberty accelerating pheremone from male mouse urine. Biol. Reprod. 15: 260-265.

Velich, A. and Ziff, E. (1984). Repression of activators. Nature 312: 594-595.

Walker, M.D., Edlund, T., Boulet, A.M. and Rutter, W.J. (1983). Cell-specific expression controlled by the 5'-flanking region of insulin and chymotrypsin genes. Nature 306: 557-561.

Wasylyk, B., Wasylyk, C., Augereau, P. and Chambon, P. (1983). The SV40 72bp repeats preferentially potentiates transcription starting from proximal natural or substitute promoter elements. Cell 32: 503-514.

Wasylyk, B., Wasylyk, C. and Chambon, P. (1984). Short and long range activation by the SV40 enhancer. Nucl. Acid Res. 12: 5589-5609.

Weaver, R.F. and Weissmann, C. (1979). Mapping of RNA by a modification of the Berk-Sharp procedure: the 5' termini of 15S beta-globin mRNA precursor and mature 10S beta-globin mRNA have identical map coordinates. Nucleic Acids Res. 7: 1175-1193.

Weber, F. and Schaffner, W. (1985). Simian virus 40 enhancer increases RNA polymerase density within the linked gene. Nature 315: 75-77.

Weiher, H., Konig, M., and Gruss, P. (1983). Multiple point mutations affecting the simian virus 40 enhancer. Science 219: 629-631.

Weiss, E.H., Mellor, A., Golden, L., Fahrner, K., Simpson, E., Hurst, J. and Flavell, R.A. (1983). The structure of a mutant H-2 gene suggests that the generation of polymorphism in H-2 genes may occur by gene conversion-like events. Nature 301: 671-674.

Winter, G. and Fields, S. (1980). Cloning of influenza cDNA into M13: the sequence of the RNA segment encoding the A/PR/8/34 matrix protein. Nucleic Acid Res. 8: 1965-1974.

Winter, G., Fields, S. and Rattig, A. (1981). The structure of two subgenomic RNAs from human influenza virus. Nucleic acids Res. 9: 6907-6915.

Wu, C. (1984). Activating protein factor binds *in vitro* to upstream control sequences in heat-shock gene chromation. Nature 311: 81-84.

Young, R.A., Hagenbuchle, O., and Schibler, U. (1981). A single mouse alpha-amylase gene specifies two different tissue-specific mRNAs. Cell 23: 451-458.

Ziff, E.B. and Evans, R.M. (1978). Coincidence of the promoter and capped 5' terminus of RNA from the adenovirus 2 major late transcription unit. Cell 15: 1463-1475.

Abbreviations used in text

A	adenine
Ad	adenovirus
AIDS	autoimmune deficiency syndrome
ATP	adenosine 5' triphosphate
bp	base pair
BHK	baby hamster kidney cells
C	cytidine
cAMP	cyclic (3' - 5') adenosine monophosphate
cDNA	DNA copy of RNA
cpm	counts per minute
CsCl	caesium chloride
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
DEAE	diethylaminoethyl
DHFR	dihydrofolate reductase
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DTT	dithiothreotol
E	enhancer core sequence
EDTA	diaminoethanetetra-acetic acid
G	guanine
GH	growth hormone
GRE	glucocorticoid responsive element
HCl	hydrogen chloride
HMG CoA	3-hydroxy-3-methylglutaryl coenzyme A reductase
HSE	heat-shock element

HSV	herpes simplex virus
IEF	isoelectric focusing
K	C or T
kb	kilobase pair
LTR	long terminal repeat
M	A or C
MHC	major histocompatibility complex
MMTV	mouse mammary tumor virus
MRE	heavy-metal responsive element
mRNA	messenger RNA
MT	metallothionein
MUP	major urinary protein
NF1	nuclear factor 1
PEG	polyethylene glycol
pH	-log [H+]
PPO	2,5 - diphenyloxazole
POPOP	1,4-bis-2-(4-methyl-5-phenyloxazolyl)-benzine
poly(A)+	polyadenylic acid
poly(A)+RNA	polyadenylated RNA
R	A or G
rev)	reverse sequence
RNA	ribonucleic acid
RNase	ribonuclease
S	C or G
S1	single strand specific nuclease
Sp1	specific transcription factor 1
SV40	simian virus 40
T	thymidine

T4	thyroxine
TMP	thymidine monophosphate
TCA	trichloroacetic acid
Tris	tris-[hydroxymethyl]-aminomethane
W	A or T
w/v	weight per volume
Y	C or T

Evolutionary amplification of a pseudogene

(mouse major urinary protein/*Mup* genes/nonsense mutation/gene family)

P. GHAZAL, A. JOHN CLARK, AND JOHN O. BISHOP

Department of Genetics, University of Edinburgh, Edinburgh EH9 3JN, Scotland

Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4182–4185, June 1985
Genetics

Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4182–4185, June 1985
Genetics

Evolutionary amplification of a pseudogene

(mouse major urinary protein/*Mup* genes/nonsense mutation/gene family)

P. GHAZAL, A. JOHN CLARK, AND JOHN O. BISHOP

Department of Genetics, University of Edinburgh, Edinburgh EH9 3JN, Scotland

Communicated by Alan Robertson, February 20, 1985

ABSTRACT The family of mouse major urinary protein (MUP) genes has about 35 members, clustered together on chromosome 4. Most of the genes belong to two major subfamilies (group 1 and group 2) each with 12–15 members. Recently we showed that most of the group 1 and group 2 genes are arranged in pairs, each containing a group 1 and a group 2 gene in divergent transcriptional orientation, with 15 kilobases of DNA between the two cap sites. Here we present the nucleotide sequence of the first exon of six group 1 genes and four group 2 genes. The data confirm the close relationship of the genes within each group and the considerable divergence of the two groups from each other. The four group 2 genes all carry the same nonsense mutation in codon 7 of the sequence that specifies the mature protein. Thus, not only do these genes have a common ancestor, but also it seems that their amplification followed the mutation of the ancestor to a pseudogene. Taking into account the 3′ flanking regions of the two genes, the overall size of each gene-pair is about 45 kilobases. The sequencing data supports our earlier suggestion that this 45 kilobase domain is the unit of *Mup* amplification.

The mouse major urinary protein (MUP) is a family of closely related polypeptides that are synthesized and secreted by the liver and excreted in the urine (1, 2). MUP mRNA makes up about 5% by weight of male liver mRNA (3, 4). Smaller amounts of biologically active mRNA are found in the lachrymal, salivary, and mammary glands. *In vitro* translation of hybrid-selected MUP mRNA from the different tissues shows that each directs the synthesis of a different subset of MUP polypeptides (5). The level of MUP mRNA in the liver is influenced by insulin, growth hormone, thyroxine, and testosterone (6). *In vitro* translation of mRNA from livers taken from mice maintained under different hormonal regimes shows that different species of mRNA (directing the synthesis of different polypeptides) respond differently to the various hormones. Testosterone is known to increase the rate of synthesis of MUP mRNA (7). The mouse genome contains about 35 MUP genes, defined as sequences that hybridize with MUP-specific probes. Most of these can be assigned to two main groups, group 1 and group 2, by hybridization with two canonical group 1 and group 2 probes (8). Most of the group 1 and group 2 genes are arranged in head-to-head (divergently orientated) pairs (9). Each pair contains a group 1 and a group 2 gene, homologous 5′ flanking sequences (two of 5 kb), 3′ flanking sequences (two of 11 kb) that contain regions of homology interspersed with nonhomologous regions, and 6 kb of DNA (located between the homologous 5′ flanking sequences) that is not duplicated within the pair. The overall size of the head-to-head pair, from the far end of one 3′ flanking sequence to the far end of the other, is about 45 kb. We have argued that this is the principal unit of MUP gene organization and evolution (9). Here we show that four group 2 genes are pseudogenes, in the sense that they contain at least one stop codon in the MUP reading-frame (the reading-frame of the group 1 genes). All of these genes contain the same stop codon in exon 1, showing that they are derived from a common ancestral pseudogene.

MATERIALS AND METHODS

The MUP genes studied here were isolated from genomic clones that have been described (8–10). Plasmid subclones and M13 mp8 and M13 mp9 subclones were isolated by standard methods and sequenced as described (11).

RESULTS AND DISCUSSION

Four Group 2 MUP Genes Are Pseudogenes. The structure of the 45-kb gene pair is shown in Fig. 1*A*. Fig. 1*B* shows the seven-exon structure (12) of the group 1 MUP genes. The nucleotide sequences of the first exon of nine different MUP genes are summarized in Fig. 2. All of these were isolated from nuclear DNA of inbred BALB/c mice and, therefore, are different members of the gene family rather than allelic variants. They are all known to be different genes either because their sequences differ or because the genes themselves or their flanking regions contain different restriction enzyme recognition sites or because of deletions or insertions in their flanking sequences (8–10). Three of the nine genes were taken from clones (BS102-2, BS109-1, and BS109-2) that contain the central portion of a 45-kb gene pair, including the 5′ end of a group 1 gene and its 5′ flanking sequence and the 5′ end of a group 2 gene and its 5′ flanking sequence (see Fig. 1). Thus, these genes are definitely known to be part of the predominant 45-kb gene-pair organization (9). The other six genes are presumed also to be derived from 45-kb gene-pair units on the basis of restriction site homologies in their 5′-flanking regions.

Four of the five group 1 genes (from clones BS1, BS5, BS6, and BL1) have identical exon 1 sequences. The fifth, from clone BS109-1, differs from the others in only two nucleotides, both in the leader sequence. Fig. 2 shows the sequence of the four identical group 1 genes and the deviations from that sequence found in the other genes. It is convenient to consider separately the leader sequence, the signal peptide region, and the remainder of exon 1, the region coding for the first 14 amino acids of the mature group 1 protein.

In the leader sequence, four of the five group 1 genes and three of the four group 2 genes are identical. These define group 1 and group 2 consensus sequences, which differ in 11 nucleotides (11/65 = 17%). In addition, the group 1 consensus sequence is 1 nucleotide longer than the group 2 consensus. One of the group 1 genes, from clone BS109-1, differs from the consensus in two positions. Similarly, one group 2 gene, from clone BL25, differs from the group 2 consensus in four positions and is the same length as the group 1 leader sequence, rather than 1 nucleotide shorter.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: MUP, mouse major urinary protein; kb, kilobases.

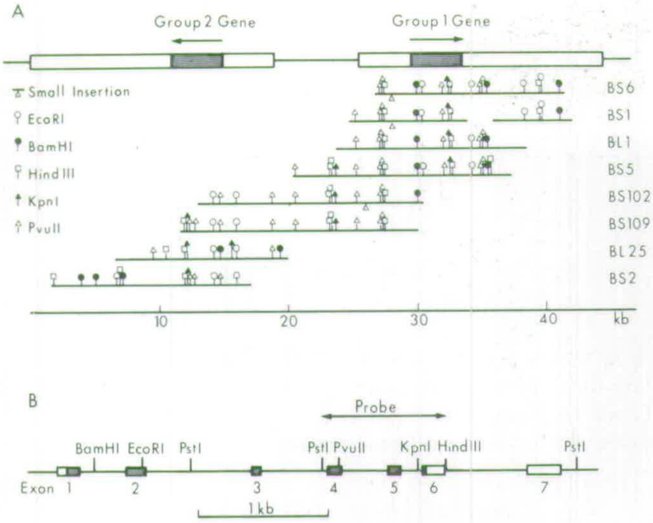


FIG. 1. Organization and structure of MUP genes. (*A*) The top line is a diagrammatic representation of the 45-kb unit. The group 1 and group 2 genes are shaded. Open rectangles are regions of homology between the flanking regions of group 1 and group 2 genes. The homology is not continuous over the 11-kb 3′ flanking region but rather is interrupted by regions of nonhomology. Restriction site maps of the genomic MUP clones used in this study are aligned with the diagram. Three small insertions and a 1.9-kb deletion are proposed in order to maximize the degree of restriction site homology between the clones. The isolation of these clones is described in refs. 8–10. (*B*) Structure of a group 1 MUP gene (12). Exons are shown as boxes, and introns, as lines. The coding region is shaded. The region of the homologous canonical group 1 and group 2 probes is shown.

The nucleotide sequence of the signal peptide region is identical in all five group 1 genes and specifies a signal peptide 18 amino acids long. In contrast, the signal peptide regions of the four group 2 genes (defined as the sequence from ATG to the NH_2 terminus of the mature group 1 protein) are all different. The signal peptides that they specify vary in length from 19 (clone BL25) to 25 (clone BS102-2) amino acids. Most of the additional codons are CTG (leucine) codons that may have arisen from adjacent CTG codons by polymerase "slippage" during replication or by unequal crossing-over. To either side of the additional codons, two of the four group 2 genes are identical and differ from the group 1 genes at two positions. BS109-2 and BL25 contain further nucleotide differences in the signal peptide region.

In the third region of exon 1, which corresponds to the NH_2-terminal 14 amino acids of the mature group 1 proteins, the group 1 genes are again identical. The group 2 genes show a clear consensus, which differs by five nucleotides from the group 1 sequence (5/42 = 12%). Clones BS102-2 and BL25 each differ from the group 2 consensus in one position in this region.

One of the differences between the group 1 and group 2 consensus is between a glycine (GGA) in the group 1 sequences and a stop codon (TGA) in the group 2 sequences (Fig. 2, amino acid 7, position 160). Thus, in the context of the group 1 genes, all four group 2 genes are pseudogenes and contain an identical lesion. Other lesions are also present. BL25 contains a stop codon in place of amino acid 2 of the mature protein (Fig. 2, position 145), and BS2 contains a second stop codon and a frameshift mutation (unpublished data). However, the stop codon that is common to the group 2 genes is their most significant feature, implying as it does that it was present in an ancestral gene, which was therefore also a pseudogene and was ancestral to all four group 2 genes shown in Fig. 2.

We identify group 1 and group 2 genes on the basis of their hybridization with two homologous genomic probes (8) that contain exons 4, 5, and 6 (Fig. 1*B*). So far we have isolated only four group 2 genes that contain exon 1. Since all of these contain the common stop codon, it is likely that all of the approximately 12 group 2 genes in the BALB/c genome share this lesion and are descended from the same ancestral pseudogene.

Evolutionary Divergence of Group 1 and Group 2 Genes. The complete nucleotide sequences of a group 1 gene (clone BS6) and a group 2 gene (clone BS2) have been determined (unpublished data). The coding regions (excluding the signal peptide region) have been identified and compared (13) with each other and with a homologous rat α_{2u}-globulin gene (14–16). The replacement site divergence of the group 1 (clone BS6) and group 2 (clone BS2) mouse genes is ≈10% (BS6 × BS2 = 10.3%) while the divergence of each mouse gene from the rat α_{2u}-globulin (clone 207) gene is ≈20% (BS6 × 207 = 19.1%; BS2 × 207 = 22.4%).

The evolution of a multigene family is more complex than the evolution of a unique gene. In the latter case, the divergence time of two contemporary genes in different species can be taken to be the time since the divergence of the two phylogenetic lines from their common ancestor. In the case of a multigene family, genes that already have diverged from each other coexist within the same genome. The contemporary MUP genes show many examples of this, with divergences that vary from 1% (different group 1 genes) to 10% (group 1 genes compared with group 2 genes). Thus, extrapolating backwards in time, it is quite possible that genes ancestral to the group 1 genes, the group 2 genes, and the rat genes had already diverged from each other in the common ancestor of rats and mice.

The members of a multigene family do not necessarily diverge within a species at rates comparable to the divergence of single genes between species. Indeed, there is strong evidence to the contrary in the present case. A set of rat α_{2u}-globulin cDNA clones, which presumably represent the more abundantly transcribed genes of the rat multigene family, are all identical in sequence (14) and very similar to the corresponding regions of a gene (15). Similarly, the abundantly transcribed group 1 MUP genes are very closely related. The rat genes and the group 1 MUP genes must have arisen from a common ancestral gene, and yet they differ by about 20% in nucleotide sequence, while at the same time different rat genes differ from each other by only 1–2% and different group 1 MUP genes differ from each other to a similar extent. The group 2 MUP pseudogenes are a third reasonably homogeneous group of genes that differ from the rat genes by about 20% and from the group 1 MUP genes by about 10%.

The explanation of this phenomenon presumably relates to the clustering (17, 18) of both the group 1 and the group 2 MUP genes (8, 9) on mouse chromosome 4. One possibility is that the ancestor of rats and mice contained rather few urinary protein genes and that different members of that small set of genes were separately amplified, by tandem duplication, in the rat and mouse lines. According to this view, the group 1 and group 2 MUP genes would have been amplified together within the 45-kb unit of genomic organization. If the common group 2 nonsense mutation arose prior to or early in the course of this amplification, it could have been carried passively, so to speak, through the amplification process, in effect, as an inert DNA sequence within the 45-kb unit.

However, it is unlikely that separate amplification processes occurred independently in the rat and mouse lines. It seems more probable that the genes were already amplified in the common ancestor. If so, what we have to explain is an apparently concerted evolution of evolutionarily diverging arrays of genes in each of the two lines. One unavoidable implication of this model is that the ancestral gene array must have been lost or replaced in one or both of the descendant

```
          Leader sequence
          Cap site
                   10        20        30        40        50        60
G1-CON    GGAGTGTAGCCACGATCACAAGAAAGACGTGGTCCTGACAGACAGACAATCCTATTCCCTACCAAA
G1-109                G               T
          ------------------------------------------------------------------
G2-1      A        G  AC     C                    C     T   T     T AG     -
G2-2      A        G  AC     C                    C     T   T     T AG     -
G2-3      A        G  AC     C                    C     T   T     T AG     -
G2-4               G  AC     C        T           C     T   T     T AG AA


          Signal peptide
              70           80           90            100          110
G1-CON    ATG AAG --- --- --- --- --- --- --- ATG CTG CTG CTG CTG TGT TTG
          ---------------------------------------------------------------
G2-1              CAG --- CTG CTG CTG CTG CTG C         C
G2-2              CAG CAG CTG CTG CTG CTG CTG C         C
G2-3              CAG CAG --- --- CTG CTG CTG C         C
G2-4            A CCA --- --- --- --- --- --- C            G


                           End signal peptide
                120           130          140            150
G1-CON    GGA CTG ACC CTA GTC TGT GTC CAT GCA   GAA GAA GCT AGT TCT ACG
          -----------------------------------   -----------------------
G2-1            A                                             G      T
G2-2            A                                             G      T
G2-3            A       T     C                               G      T
G2-4       A    A  T                                T         G      T


                               End exon 1        Number of nucleotides
          160          170          180      Leader Signal Remainder Total
G1-CON    GGA AGG AAC TTT AAT GTA GAA AAG      66     54       42     162
          -------------------------------    -----------------------------
G2-1      T                   A         A      65     72       42     179
G2-2      T        C          A         A      65     75       42     182
G2-3      T                   A         A      65     69       42     176
G2-4      T                   A         A      66     57       42     165
```

FIG. 2. MUP gene exon 1 sequences. The complete sequence of exon 1 of nine MUP genes is shown. Above the dashed line, five group 1 (G1) genes are shown: Four of these, clones BS1, BS5, BS6, and BL1, are identical in exon 1 and are shown as the sequence labeled G1-CON (for consensus). The fifth, G1-109, differs from the consensus in only two positions. Below the dashed line, the differences between four group 2 (G2) genes and the group 1 consensus are shown. G2-1, G2-2, G2-3, and G2-4 are, respectively, clones BS2, BS102-2, BS109-2, and BL25. The absence of a nucleotide relative to other sequences is signified by a dash. The G→T stop-codon mutation (nucleotide 160) is underlined. The cap site was defined by S1 nuclease mapping and primer extension (unpublished data).

lines. The contemporary arrays would have been developing at the same time. As in the case of the simpler model, the principal unit of MUP gene evolution would be the 45-kb gene pair.

The urinary protein genes of rats and mice invite comparison with the rDNA of *Xenopus laevis* and *X. borealis*. The spacer sequences of the tandemly arranged rDNA genes have diverged widely between the two species, but within each species they are relatively homogeneous (19). This has been explained by a model incorporating two main features: unequal sister-strand crossing-over within the tandem arrays and selective constraints on their size (20). Under these circumstances it can be shown that the entire contemporary array in a given species may be directly descended from a single member of the array at some past time. Thus, genetic drift can go hand-in-hand with the preservation of homogeneity within the array. The degree of homogeneity preserved will depend on the mutation rate, frequency of unequal crossing-over, selection pressure, and so on (21). At least some of the 45-kb MUP gene pairs are arranged tandemly and in direct orientation (9). If this arrangement were general, the unequal crossing-over model would provide a sufficient explanation for the replacement of the ancestral array with a new one.

The phenomenon also can be explained on the basis of gene conversion (22, 23). Unequal crossing-over and gene conversion were discussed previously in relation to MUP gene evolution (9). Ohta (21) has shown that unequal crossing-over and gene conversion can provide formally equivalent explanations of the concerted evolution of a gene family.

We have suggested two models to explain the contemporary relationships of the rat genes and the group 1 and group 2 MUP genes: (*i*) separate *de novo* amplification of different genes in the rat and mouse lines and (*ii*) the replacement of a preexisting array by a new one in each line by unequal crossing-over, gene conversion, or both. The idea central to both models, that the unit of MUP gene amplification is the 45-kb gene pair, is suggested by the 10–15 copies of the gene pair that are present in the genome of the laboratory (BALB/c) mouse (9). The same idea also can explain the divergence of the group 1 and group 2 genes during a time period in which each group remained reasonably homogeneous. According to the models, the two main parts of the unit, the group 1 and group 2 genes and their respective flanking sequences, cannot replace each other and so would have been able to diverge. On the other hand, the 45-kb unit as a whole replaces other 45-kb units, and it is to this that we would attribute much of the uniformity of the 45-kb units and, in particular, the uniformity of the group 1 and group 2 genes themselves. Thus, we would date the onset of divergence of the group 1 and group 2 genes to a time close to that at which the 45-kb unit originated (presumably by an inversion), whether in the mouse line or in a line ancestral to the divergence of mice and rats.

The group 1 genes are more homogeneous than the group 2 genes (ref. 8; Fig. 2). This can most easily be explained by supposing that selective constraints are superimposed on the amplification–replacement process. Selection acting against an unfavorable newly arisen group 1 gene would tend to lead to the elimination of the 45-kb unit within which it was located. Similarly, selection may have maintained the homogeneity of the group 2 genes up to the time at which the pseudogene mutation occurred but presumably did not operate on the pseudogene and its descendants. If so, mutational changes in the group 2 pseudogenes would have accumulated more rapidly. This raises the question as to why the group 2 pseudogene mutation was tolerated in the first instance. The most satisfactory explanation is that the inversion and the pseudogene mutation arose at about the same time. If this is the case, we can view (*i*) the homogeneity of the group 2 genes as a function of the concerted evolution of 45-kb units, driven by the group 1 genes, and (*ii*) the inhomogeneity of the group 2 genes as a function of the underlying mutation rate, acting against the homogenization process but unaffected by selection.

T. Ohta writes (personal communication):

> The theory of concerted evolution is already available and is applicable to the present data. Referring to Ohta (21), let us assume the following parameter values: n (no. of amplification units) = 10, N (effective population size) = 10^4, v (mutation rate of nucleotides per generation) = 10^{-9}, β (interchromosomal recombination rate between units) = $10^{-4} \sim 10^{-6}$, and λ (rate by which a unit is replaced by another unit, or the rate of one cycle of unequal crossing-over or duplication–deletion) = 10^{-6}. Then the average divergence between the nonallelic genes belonging to the family becomes about 1%. By using the same set of parameter values except that the mutation rate (v) is five times as large (5×10^{-9}), one gets an average divergence of about 4.5%. The former is appropriate for group I, and the latter, for group II genes.
>
> The above application has several implications. (*i*) Even if tentative, the effective rate of unequal crossing-over is estimated. (*ii*) In the above model, nucleotide substitution is assumed to be selectively neutral. In view of available data, most nucleotide substitutions are neutral [Kimura (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, England], and the present case is not likely to be an exception. Group II genes are free to change and the rate is high. (*iii*) The time for spreading of a unit is estimated with the above set of parameters to be about 10^7 generations (see Ohta, *Genet. Res.* **41,** 47–55).

This work was supported by the Medical Research Council and the Cancer Research Campaign.

1. Rümke, Ph. & Thung, P. J. (1964) *Acta Endocrinol.* **47,** 156–164.
2. Finlayson, J. S., Asofsky, R., Potter, M. & Runner, C. C. (1965) *Science* **149,** 981–982.
3. Hastie, N. & Held, W. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1217–1221.
4. Clissold, P. M. & Bishop, J. O. (1981) *Gene* **15,** 225–235.
5. Shaw, P. H., Held, W. & Hastie, N. D. (1983) *Cell* **32,** 755–761.
6. Knopf, J. L., Gallagher, J. R. & Held, W. A. (1983) *Mol. Cell. Biol.* **3,** 2232–2240.
7. Derman, E. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 5425–5429.
8. Bishop, J. O., Clark, A. J., Clissold, P. M., Hainey, S. & Francke, U. (1982) *EMBO J.* **1,** 615–620.
9. Clark, A. J., Hickman, J. & Bishop, J. O. (1984) *EMBO J.* **3,** 2055–2064.
10. Clark, A. J., Clissold, P. M. & Bishop, J. O. (1982) *Gene* **18,** 221–230.
11. Anderson, S., Gait, M., Mayol, L. & Young, I. G. (1980) *Nucleic Acids Res.* **8,** 1731–1745.
12. Clark, A. J., Clissold, P. M., Al Shawi, R., Beattie, P. & Bishop, J. O. (1984) *EMBO J.* **3,** 1045–1052.
13. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20,** 555–565.
14. Unterman, R. D., Lynch, K. R., Nakhasi, H. L., Dolan, K. P., Hamilton, J. W., Cohn, D. V. & Feigelson, P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 3478–3482.
15. Dolan, K. P., Unterman, R., McLaughlin, M., Nakhasi, H. L., Lynch, K. R. & Feigelson, P. (1982) *J. Biol. Chem.* **257,** 13527–13543.
16. Laperche, Y., Lynch, K. R., Dolan, K. P. & Feigelson, P. (1983) *Cell* **32,** 453–460.
17. Bennett, K., Lalley, P., Barth, R. & Hastie, N. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 1220–1224.
18. Krauter, K., Leinwald, L., D'Eustachio, P., Ruddle, F. & Darnell, J. (1982) *J. Cell Biol.* **94,** 414–417.
19. Brown, D. D., Wensink, P. C. & Jordan, E. (1972) *J. Mol. Biol.* **63,** 57–73.
20. Smith, G. P. (1973) *Cold Spring Harbor Symp. Quant. Biol.* **38,** 507–513.
21. Ohta, T. (1983) *Theor. Pop. Biol.* **23,** 216–240.
22. Baltimore, D. (1981) *Cell* **24,** 592–594.
23. Dover, G. & Coen, E. S. (1981) *Nature (London)* **290,** 731–732.

Sequence structures of a mouse major urinary protein gene and pseudogene compared

A.J.Clark[1], P.Ghazal, R.W.Bingham[2], D.Barrett[3] and J.O.Bishop

Department of Genetics, and [2]Department of Veterinary Pathology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JN, UK, and [3]Department of Biological Sciences, University of Denver, Denver, CO, USA

[1]Present address: A.F.R.C. Animal Breeding Research Organisation, West Mains Road, Edinburgh, UK

Communicated by J.O.Bishop

Laboratory mouse strains carry ~35 major urinary protein (MUP) genes per haploid genome, tightly clustered together on chromosome 4. Most belong to two main groups (Groups 1 and 2). The available evidence strongly suggests that the Group 1 genes are active while the Group 2 genes are pseudogenes. Here we present the complete sequence of a Group 1 gene and a Group 2 gene and 700 bp of flanking sequence. The sequence of the Group 1 gene is consistent with its being active. The Group 2 gene contains two stop codons and a frame-shift mutation in the reading frame defined by the Group 1 gene, and would code for a signal peptide 25 rather than 19 amino acids long. The Group 2 gene differs from the Group 1 gene in other ways: a deletion upstream of the TATA box and another in intron 3, a base change in the TATA box itself, a 2 bp duplication at the splice acceptor boundary of intron 6, an altered poly(A) addition signal and a 1-base deletion 5′ to the initiation codon. Some of these differences may explain the 10- to 20-fold higher level of Group 1 mRNA in mouse liver, and the fact that Group 1 and Group 2 transcripts are mainly spliced differently. The presence of the stop codon means that the Group 2 gene is a pseudogene in the context of the Group 1 gene. However, there is some evidence that the mature hexapeptide that it would code for may have biological activity. The 12 acceptor splice sites of the two genes all contain the identical sequence ACAG at the exon boundary. As a result this region shows an unusually high level of base-pairing homology with the splice donor site. A sequence showing a moderate to high homology with the sequence CTGAC is found between 17 and 35 bp 5′ to the acceptor site boundary in every intron.

Key words: mouse/major urinary protein/pseudogene/sequence/comparison

Introduction

The mouse major urinary proteins (MUPs) are a closely related group of small acidic proteins which are synthesised in the liver, secreted into the blood and subsequently excreted in the urine. There are ~35 MUP genes in the mouse genome (Bishop *et al.*, 1982). On the basis of nucleic acid hybridisation experiments the 35 genes can be subdivided into two groups (Group 1 and Group 2), each with ~15 members, and a small number of other genes not closely related to either group. The Group 1 and Group 2 genes are part of large units of DNA organisation which are ~45 kb long (Clark *et al.*, 1984b; Bishop *et al.*, 1985). Each unit contains one Group 1 gene and one Group 2 gene, ~15 kb apart, in a divergent transcriptional orientation (i.e., head-to-head organisation). Here we present the full sequence of the transcription units of a Group 1 and a Group 2 gene, and also some 700 bp of flanking sequence. We show that the Group 2 gene, with two stop codons and a frame-shift mutation, is a pseudogene in the context of the Group 1 gene. However, we cite evidence that raises the possibility that the hypothetical oligopeptide product of the Group 2 gene may have biological activity. Several other differences between the Group 1 and Group 2 genes were observed, some of which may impair the efficiency of transcription or translation of the latter.

Results

Figure 1A shows the basic arrangement of Group 1 and Group 2 genes and the regions of DNA sequenced. Figure 1B and C shows M13 clones that were generated, respectively, from BS6 (Group 1) and BS2,3 and sequenced. BS2,3 is the name given to a Group 2 gene which, with its flanking regions, is defined by two overlapping clones. In the case of BS6, 568 bp of 5′-flanking sequence, the 3917 bp transcription unit and 136 bp of 3′ flanking sequence were determined. Approximately 80% of the sequence was determined on both strands. The region of BS2,3 homologous to that determined for BS6 was sequenced primarily on one strand.

Determination of the Group 1 mRNA cap site

We previously described the sequence of the combined exons of BS6. The gene encodes a short mRNA of ~750 nucleotides within six exons and a long mRNA of 882 nucleotides within seven exons (Clark *et al.*, 1984a). The two forms are generated by different splicing events. The long mRNA is considerably more abundant. Previously we positioned the mRNA cap site provisionally. On the basis of two criteria, S1 nuclease protection and primer extension, we now confirm that it is located 30 ± 1 bp downstream from the TATA box (Figure 2).

Comparison of BS6 and BS2,3

Figure 3 shows the sequences of BS6 and BS2,3, aligned to maximise base-pairing homology between them. The boxes surround the exons previously defined for BS6 (Clark *et al.*, 1984a).

Insertions and deletions. The comparison shows that there are three large insertions or deletions (>17 bp) and 20 smaller insertions or deletions (<9 bp). Otherwise the two sequences are co-linear over the entire sequenced region. The most 5′ large insertion or deletion occurs within a very A-rich tract located 50 bp 5′ to the start of each transcription unit. In BS6 this tract (primarily A, occasionally interrupted by C) is 44 bp long, whereas it is only 16 bp long in BS2,3. To date, the corresponding regions of nine different MUP genes (five Group 1 and four Group 2) have been sequenced. Many show variation in the length of the A-rich tract, from a minimum of 11 bp to a maximum of 61 bp (P.Ghazal, unpublished observations). The second major interruption in the co-linearity of the two sequences occurs in the first

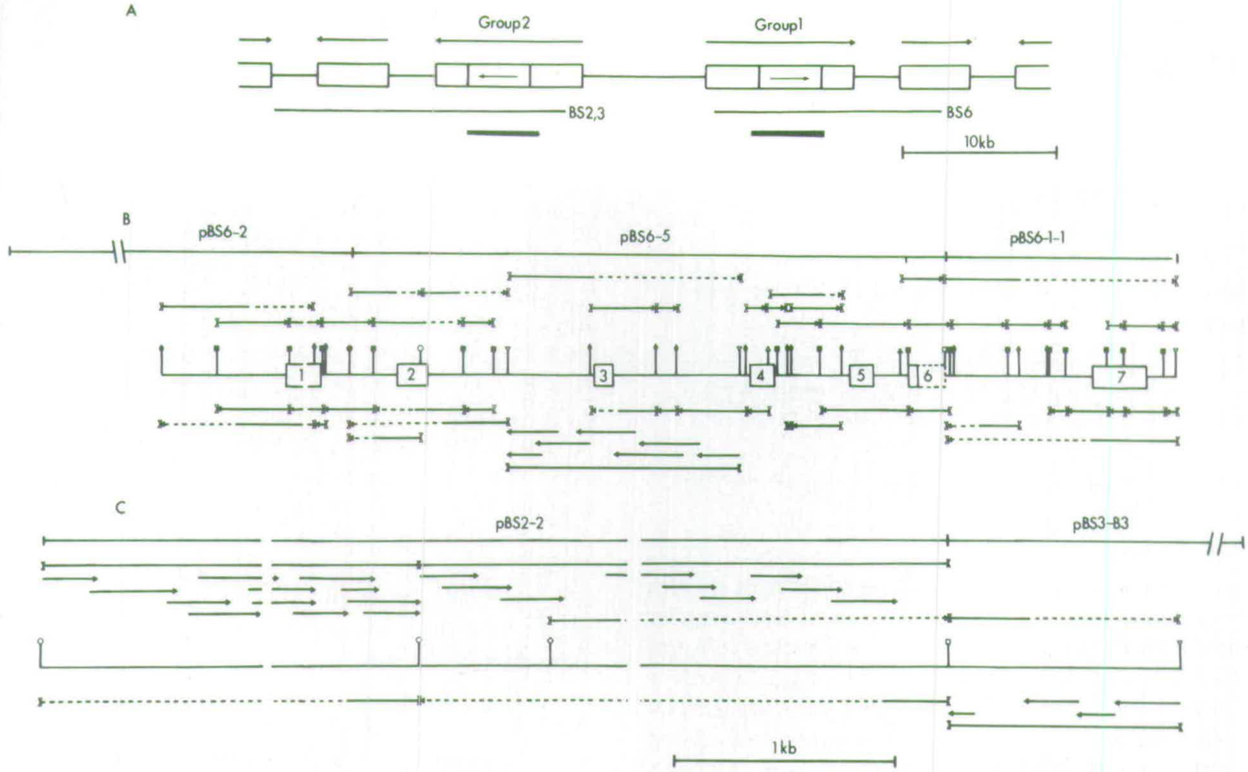


Fig. 1. Sequencing strategy for BS6 and BS2,3. **A:** The predominant arrangement of Group 1 and Group 2 genes and their flanking sequences in the BALB/c genome. Regions of inverted symmetry are shown as boxes with arrows above them. The Group 1 and Group 2 transcription units are marked as boxes containing arrows which indicate the direction of transcription. The continuous lines below show the relationship of the lambda clones to the chromosome map. BS2,3 is a composite of two Group 2 lambda clones which overlap extensively and have identical restriction enzyme sites in this region of overlap. — Indicates the regions that were sequenced. **B:** Sequencing strategy for BS6. ⊢⊣, the plasmid subclones from which M13 clones were derived.]—[, M13 clones which were cloned at specific sites: continuous line, region sequenced; broken line, remainder of the clone which was not sequenced. →, M13 clones for which the RF was prepared and the insert progressively shortened by the method of Hong (1982). Arrows indicate the regions sequenced. Arrowheads show the direction of sequencing. The restriction map covers the region sequenced and shows the sites employed for the M13 cloning; ●, *Bam*HI; ○, *Eco*RI; □, *Hind*III; ▲, *Kpn*I; △, *Pvu*II; ▽, *Pst*I; ◇, *Aha*III; ■, *Sau*IIIA; ◆, *Alu*I. The numbered, open boxes show the positions of the exons, and the dashed extension of exon 6 shows the position of those sequences that are present in short MUP mRNA. **C:** Sequencing strategy for BS2,3. Symbols are the same as in **B**. The scale is the same for **B** and **C**.

exon within the region which codes for the signal peptide. BS6 has a 19 amino acid and BS2,3 a 25 amino acid signal peptide, the difference being due to a net insertion of six leucine residues (6 × CGT) in BS2,3. The length of this region is different in each of four Group 2 MUP genes. In contrast, the sequences of the entire signal peptide region of five Group 1 genes are identical (Ghazal *et al.*, 1985). The third major insertion or deletion is in the third intron and occurs in a region of DNA that consists primarily of runs of GT and GTT. In BS6 this region (+1537 to +1633) is 97 bp long, whereas the homologous region in BS2,3 (+1542 to +1557) is only 16 bp long. Comparable sequence data from other MUP genes are not available. However, restriction site mapping suggests that there are no large differences in length between different Group 1 genes or between different Group 2 genes.

Transcription initiation signals. The DNA sequence signals which are presumed to be required for transcription are listed in Table I. There is a possible 'CAAT' box at −109 in BS6 and −77 in BS2,3, although the sequences are considerably diverged from the published consensus, sharing only 5/9 positions, one of which is an unspecified pyrimidine in the consensus sequence. Both BS6 and BS2,3 have a consensus 'TATA' box at −31. BS2,3, however, contains a G at a position normally occupied by an A (Table I).

Splice sites. Table I also tabulates the donor and acceptor splice sites of the six introns of each gene. All 24 sites accord with the GT/AG rule and show a good agreement with the consensus sequences derived by Breathnach and Chambon (1981). In the six donor sites, BS6 and BS2,3 differ in a total of two positions (2/36 bp). Similarly the two genes differ by a total of two positions in five of the six acceptor sites (2/50 bp). The acceptor site in intron 6 of BS2,3 has a net insertion of 2 bp compared with BS6. The mRNA transcribed from Group 2 genes is mainly of the short variety which lacks exon 7 and contains an extended exon 6, while the mRNA transcribed from the Group 1 genes is mainly the longer variant which contains the short exon 6 spliced to exon 7 (Clark *et al.*, 1984a). It seems possible that the net insertion of 2 bp in BS2 may underly this difference by partially inactivating the acceptor site of intron 6.

Transcription termination signals. Most Group 1 MUP mRNA contain the 250 bp long untranslated exon 7. In this exon at +3895 there is a poly(A) addition signal (AATAAA). By comparison with the sequence of a number of MUP cDNA clones (Kuhn *et al.*, 1984; Clark *et al.*, 1985) this sequence is found to be located 22 bp 5′ to the beginning of the poly(A) tract. An identical poly(A) addition signal is present in the homologous position in the BS2,3 sequence. The less abundant short forms of Group 1 mRNA which terminate at the end of an extended exon 6 are polyadenylated at sites that relate to the rare poly(A) addition site ATTAAA at +2964 in the BS1 sequence and the usual AATAAA site at +2979 (Clark *et al.*, 1984a). The sequence in BS2,3 that corresponds in position to the first of these

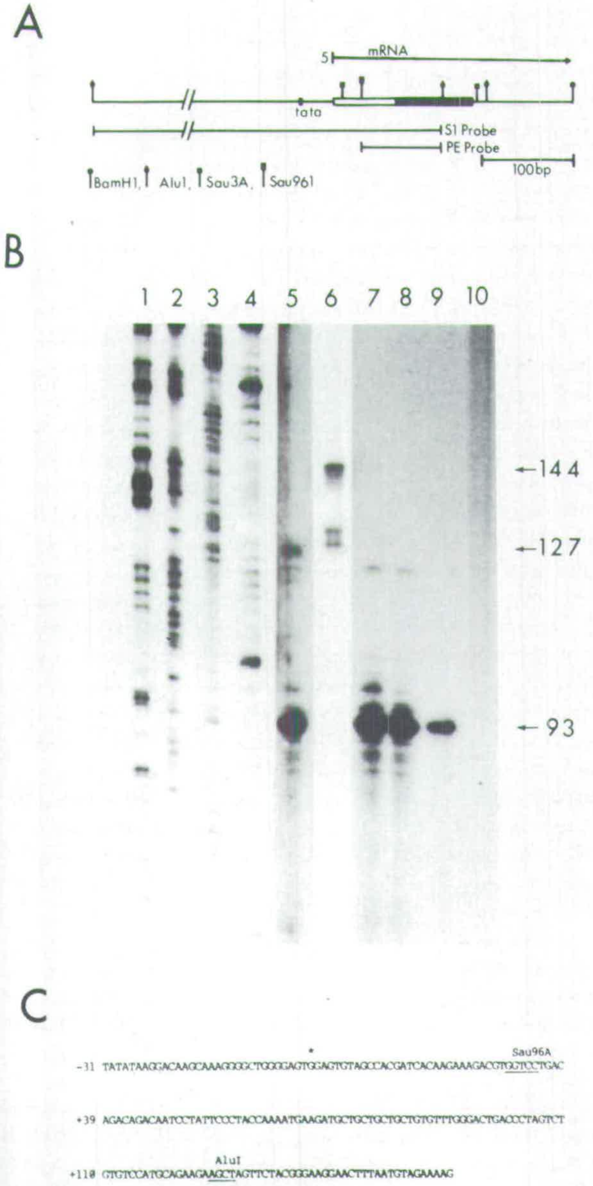

A
5
mRNA
tata
S1 Probe
PE Probe
100bp
BamH1, Alu1, Sau3A, Sau961
B
1 2 3 4 5 6 7 8 9 10
-144
-127
-93
C
Sau96A
-31 TATATAAGGACAAGCAAAGGGGCTGGGGAGTGGAGTGTAGCCACGATCACAAGAAAGACGTGGTCCTGAC
+39 AGACAGACAATCCTATTCCCTACCAAAATGAAGATGCTGCTGCTGCTGTGTTTGGGACTGACCCTAGTCT
AluI
+11Ø GTGTCCATGCAGAAGAAGCTAGTTCTACGGGAAGGAACTTTAATGTAGAAAAG

Fig. 2. S1 protection and primer extension. **A:** Restriction map of the 5′ region of MUP BS6 showing its relationship to the probes used for S1 protection and primer extension. Open and closed boxes show the untranslated and translated regions of exon 1. **B:** Electrophoretic analysis of the products of S1 protection and primer extension. **Lanes 1–4**, sequence ladder of an M13 clone used to provide mol. wt. markers. **Lane 5**, primer extension of liver poly(A)$^+$ RNA. **Lane 6**, S1 protection of liver poly(A)$^+$ RNA. **Lanes 7–9**, primer extension controls: **7**, kidney poly(A)$^+$ RNA; **8**, no RNA; **9**, primer extension probe alone. **Lane 10**, S1 protection of kidney poly(A)$^+$ RNA (S1 protection control). The primer extension probe is 93 bp long (see **C**). In the two tracks with liver poly(A)$^+$ RNA, both the S1 analysis and the primer extension analysis yield bands 127–128 bp long which positions the mRNA cap site 30 bp ± 1 bp downstream from the TATA box. An artifactual band at 144 bp is observed in **lane 6** which results from partial homology of the MUP mRNA sequences immediately 3′ to the *Alu*I sequences to the polylinker region of M13 that was present in the S1 probe. **C:** The sequence from the TATA box through the cap site (·) to beyond the *Alu*I site. The primer extension probe is the fragment from *Sau*961A to *Alu*I.

is AATAAA (+2893) and to the second GATAAG (+2907) which has not been reported to be a poly(A) addition site. There are no other AATAAA or ATTAA sequences in the region of BS2,3 within which short mRNA terminates. The present results offer a second explanation of the preponderance of short mRNA among the Group 2 transcripts: differences in the extent to which exon 7 is spliced into the mRNA may be due to differences in these 'internal' poly(A) addition signals rather than to the differences in the splice sites described above.

The coding region. The consensus sequence CCRCC has been shown to be conserved immediately 5′ to the AUG of the N-terminal methionine in a large number of eukaryote mRNAs and is proposed to be involved in ribosome binding (Kozak, 1984a). Within this consensus the R (usually A) at −3 from AUG is the most highly conserved residue, and its mutation to C in the rat pre-proinsulin gene dramatically reduced the efficiency of translation (Kozak, 1984b). The sequence immediately 5′ to ATG in BS6, CCAAA, conforms reasonably well with the consensus. In BS2,3, however, a 1 bp deletion relative to BS6 brings a C into the −3 position, thus raising a question as to whether BS2,3 transcripts would efficiently initiate translation.

Group 2 genes are transcribed much less abundantly than Group 1 genes (Clark *et al.*, 1984a). The combined exonic sequence of BS2,3 could not code for a mature MUP protein because it has stop codons in exon 1 (+156) and exon 3 (+1422) and a frame-shift mutation in exon 3 (+1472 to +1473) which generates a stop codon at +1482. In the other two frames BS2,3 contains no long open reading frames. Thus BS2,3 is a MUP pseudogene in that it has three lesions which make it untranslatable. We showed previously that three other Group 2 genes share the same stop codon in exon 1 and that the mutation therefore is probably ancestral to the Group 2 lineage (Ghazal *et al.*, 1985).

Discussion

Splice sites and intronic sequences

An interesting feature of the six acceptor sites in each of the genes is the absolute conservation of the four 3′-terminal bp. The splice acceptor site consensus sequence, derived from many genes, is NCAG, where A and G are absolutely conserved, C is present in 80% of cases, and N can be any base. In all six sites of both MUP genes this sequence is ACAG, the most notable feature being the conservation of the first A. The consensus NCAG is drawn from a large sample of different genes (Breathnach and Chambon, 1981), and would obscure such a feature of any single gene. We have therefore examined the acceptor sites of a number of genes that have multiple introns: mouse dihydrofolate reductase (Nunberg *et al.*, 1980; Crouse *et al.*, 1982; Simonsen and Levinson, 1983), alpha-fetoprotein (Law and Dugaiczyk, 1981; Eiferman *et al.*, 1981; Gerin *et al.*, 1981), alpha-amylase (Hagenbuechle *et al.*, 1981; Young *et al.*, 1981), MHC genes H-2 K-B (Weiss *et al.*, 1983) and H-2 L-D (Moore *et al.*, 1982; Evans *et al.*, 1982) and chicken alpha-2 collagen (Dickson *et al.*, 1981; Wozney *et al.*, 1981). In all cases, the terminal AG of the acceptor is absolutely conserved, but only in the case of the MUP genes is either of the two preceding nucleotides absolutely conserved.

The conserved A and C residues are complementary to the absolutely conserved G and T of the splice donor sites. We therefore asked how many base pairs would be made between five bases at the donor site of each intron (GTNNN) and the sequence NNNAC of the same intron. In nine cases three and in two cases four base pairs could be made (Table II). The probability of this arising by chance is very small (3×10^{-5}), due almost entirely to the absolute conservation of the donor site T and G residues and the acceptor site A and C residues. This highly non-random complementarity between the two regions suggests that they may come together at some stage in the splicing process. To ask

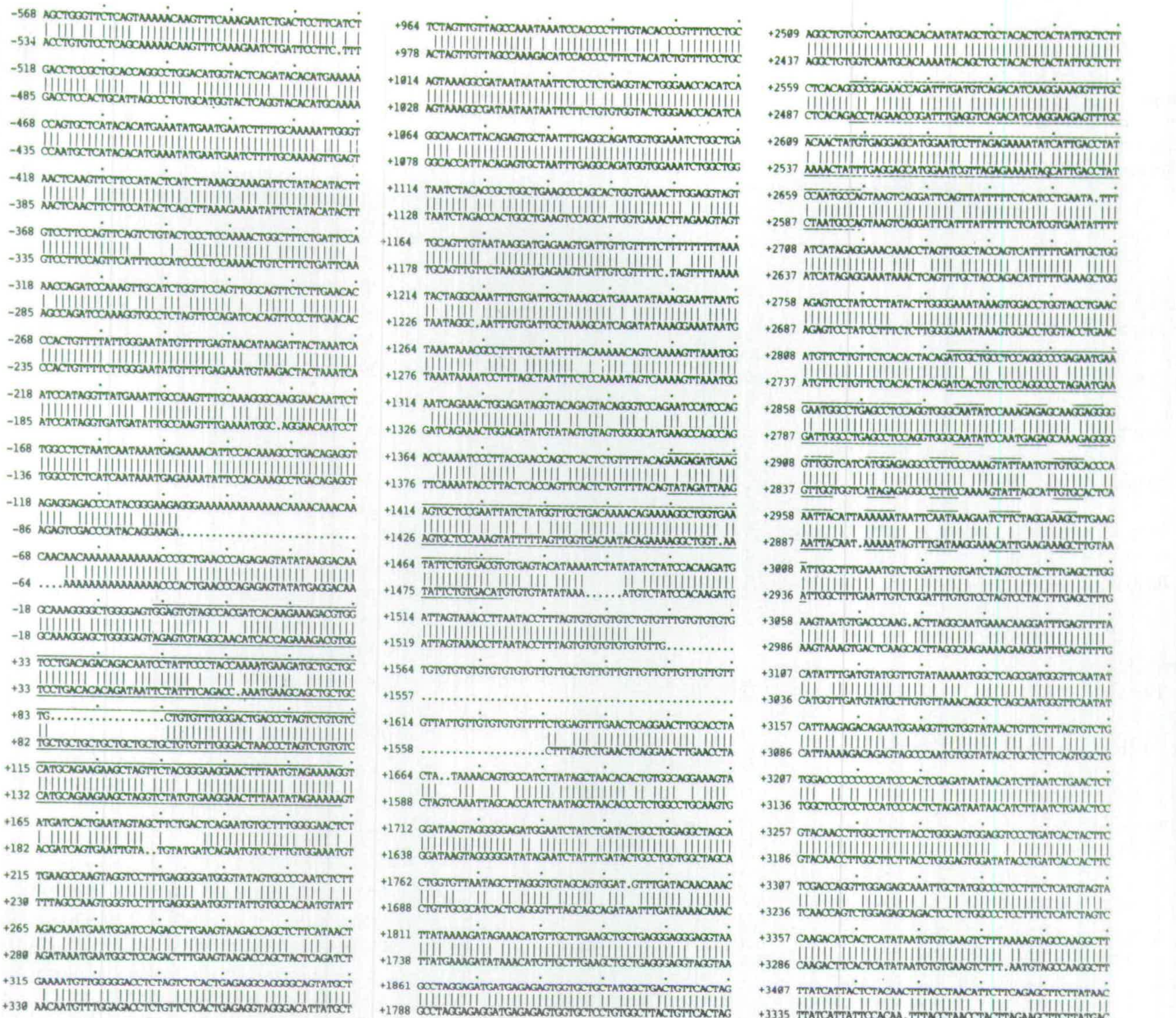


Fig. 3. Sequence comparison of BS6 and BS2,3. The sequences of BS6 (Group 1) and BS2,3 (Group 2) were aligned to maximise homology using the GAP program of Devereux and Haeberli (1984). The BS6 sequence is presented in the top line of the comparison. The regions boxed by the continuous lines show exons 1–7 of the predominant 882-bp long form of MUP mRNA (Clark *et al.*, 1984). The sequences boxed by the broken lines are those present in the shorter form of MUP mRNA. The numbers refer to the distance, in bp, from the cap site.

Table I. DNA sequence signals present in BS6 and BS2,3

Signal	Gene	Sequence	Position
Transcription initiation	BS6	GACĊCATAC	−109
	BS2	GACCCATAC	−77
	Consensus*	GGYCAATCT	−80
	BS6	GAGṪATATAAGG	−31
	BS2	GAGTATATGAGG	−31
	Consensus*	GNGTATAWAWNG	−30
Donor acceptor splice sites:			
Intron 1	BS6	ĠTATGA/TCTATTACAĠ	+163/+500
	BS2	GTACGA/TCTACTACAG	+180/+513
Intron 2	BS6	ĠTAAGT/TGTTTTACAĠ	+635/+1402
	BS2	GTAAGT/TGTTTTACAG	+648/+1414
Intron 3	BS6	ĠTGAGT/TCTTCCACAĠ	+1477/+2113
	BS2	GTGTGT/TCCTCCACAG	+1488/+2041
Intron 4	BS6	ĠTAAAG/CTTCTCACAĠ	+2225/+2565
	BS2	GTAANG/CTTCTCACAG	+2153/+2493
Intron 5	BS6	ĠTAAGT/CACACTACAĠ	+2668/+2830
	BS2	GTAAGT/CACACTACAG	+2596/+2760
Intron 6	BS6	ĠTGGGC/TGGCTTACAĠ	+2877/+3667
	BS2	GTGGGC/TGGCTTACACAĠ	+2806/+3592
	Consensus*	GTRAGT/YYYYYYXCAG	
Poly(A) addition signals:			
Exon 5	BS6	ȦTTAAA, ȦATAAA	+2964, +2979
	BS2	AATAAA, GATAAG	+2893, +2907
Exon 7	BS6	ȦATAAA	+3895
	BS2	AATAAA	+3819
	Consensus*	AATAAA	
Translation initiation	BS6	CCAAAȦTG	+67
	BS2	ACCAAATG	+66
	Consensus⁺	CCRCCATG	
Translation termination	BS6 (exon 6)	ṪGA	+2854
	BS2 (exon 1)	TGA	+156
	BS2 (exon 3)	TAA	+1422
	BS2 (exon 3)	TGA	+1482

Consensus sequences were taken from Breathnach and Chambon (1981) (*) and from Kozak (1984a) (⁺)

Table II. Base-pairing homology between the splice donor sites (GTNNN) and nucleotides −7 to −3 of the splice acceptor site (NNNAC) of the same intron

Intron		BS6		BS2,3	
1	NNNAC	ATTAC	3	ACTAC	3
		|||		|||	
	NNNTG	GTATG		GCATG	
2	NNNAC	TTTAC	4	TTTAC	4
		||||		||||	
	NNNTG	GAATG		GAATG	
3	NNNAC	TCCAC	3	TCCAC	3
		|||		|||	
	NNNTG	GAGTG		GTGTG	
4	NNNAC	CTCAC	3	CTCAC	3
		| ||		| ||	
	NNNTG	AAATG		NAATG	
5	NNNAC	ACTAC	3	ACTAC	3
		|||		|||	
	NNNTG	GAATG		GAATG	
6	NNNAC	CTTAC	3	TACAC	3
		| ||		|||	
	NNNTG	GGGTG		GGGTG	

Table III. Potential splice lariat junctions in the introns of MUP genes BS6 and BS2,3

Intron	BS6					BS2,3				
	Distance from junction		X	Y	Z	Distance from junction		X	Y	Z
1	21	CTTAA	3	4	3	21	CTTAA	3	4	3
2	24	CTTAC	4	5	5	24	CTTAC	4	5	5
3	22	CTGAG	4	3	3	22	CTGAG	4	3	2
4	17	CTCAC	4	4	3	17	CTCAC	4	4	
5	24	CTGAA	4	3	3	24	CTGAA	4	3	3
6	30	ATGAA	3	2	2	31	ATGAG	3	2	1

X, Y and Z, number of positions agreeing with CTGAC, CTTAC and with the complement to the splice donor site, respectively.

whether complementarity between these two regions of an intron is general, we examined the introns of the genes listed above and also those of the mouse metallothionein (Glanville *et al.*, 1981) and alpha (Mishioka and Leder, 1979) and beta (Konkel *et al.*, 1979) globin genes for evidence of complementarity between the first five bases of the donor site and the five bases before the AG of the acceptor site of the same intron. The average complementarity was 49% which, although less than the 68% found in the MUP introns, is also high. This is partly due to the absolute conservation of position 1 of the donor site and the 80% occupancy of position −3 of the acceptor site by C, but also to the fact that donor site positions 3−5 are nearly always purines while positions −5 to −7 of the acceptor site are nearly always pyrimidines. Thus elevated complementarity between the two regions is very common. If they do associate during splicing, this could follow the association of the donor (Mount *et al.*, 1983; Kramer *et al.*, 1984) and possibly also the acceptor sites (Lerner *et al.*, 1980; Rogers and Wall, 1980) with U1 snRNP, but would presumably precede the formation of the G5′-2′A lariat junction 20 or so bases upstream (Ruskin *et al.*, 1984).

Keller and Noon (1984) discovered the consensus CTGAC 20−55 nucleotides from the acceptor site boundary in a number of introns. During the search, the A residue was required to be present because in some cases it is known to participate in the junction point of the lariat splicing intermediate (Ruskin *et al.*, 1984). It was suggested that during splicing a transient base-pairing interaction occurs between this site and the splice donor site. We searched the MUP gene introns for three pentamer sequences, CTGAC itself, CTTAC which is the complement of the donor splice consensus, and the complement of the donor splice site of the intron under scrutiny. The most consistent results were obtained with CTGAC. In every intron, between nucleotides 17 and 35, there is a sequence that matches CTGAC in either four (eight cases) or three (four cases) positions (Table III). Overall, the match of these sites to CTGAC (73%) is greater than to CTTAC (69%) or to the different donor sites of the separate introns (60%). Given the selection of the A residue, we would expect this degree of matching, or better, to occur in random DNA once per 46 bases. We observe it once per 19 bases, which is not strikingly more frequent. It seems likely, nevertheless, that this technique identifies the A residue at the lariat junction in most if not all cases.

Group 2 genes are pseudogenes in the context of Group 1 genes

While the available evidence indicates that the Group 1 genes are true genes (see Clark *et al.*, 1985), all of the Group 2 genes so far examined are putative pseudogenes. BS2,3 carries three lesions in its protein coding sequence and could not be translated to yield a protein with the mol. wt. of MUP. Partial sequence analysis of three other Group 2 genes has shown that they all contain the same stop codon in exon 1 (Ghazal *et al.*, 1985). It is likely that all Group 2 genes in the BALB/c genome share this lesion, and are descended from the same ancestral gene. Other sequence differences between BS6 and BS2,3, most of which might affect transcription or translation, are (i) upstream and intronic deletions that may affect enhancer function, (ii) a substitution of G for A in the TATA box region that may affect the strength of the promoter, (iii) a small duplication in the splice-acceptor site of intron 6 that may favour the formation of the shorter form of mRNA, (iv) an alteration in one of the poly(A) addition signals of short form mRNA (ATTAAA→AATAAA) that also may favour the formation of the shorter mRNA, (v) an alteration in the translation initiation signal CCAAA→ACCAA that may impair the efficiency of translation and (vi) an in-frame increase in the length of the signal peptide region.

A possible function for the truncated product of the Group 2 gene

Some Group 2 genes are probably transcribed to yield a short mRNA (Clark *et al.*, 1984a) although the steady-state mRNA level is much less than that observed for Group 1 genes (<10%). If the Group 2 transcripts are translated and if the polypeptides are then processed, the products will be peptides six amino acids long with a mol. wt. of 630. Such small peptides would be rapidly excreted into the urine.

Mouse urine contains androgen regulated agents that dramatically accelerate the onset of puberty when administered to young females (Vandenbergh *et al.*, 1975). One is probably a protein, with a mol. wt. >12 000, i.e., consistent with the mol. wt. of MUP. The activity of this agent largely survives proteolysis, but becomes dialysable. The second agent has a mol. wt. of 860, and seems to be one or more of a mixture of oligopeptides (Vandenberg *et al.*, 1976). These apparently contradictory observations can be reconciled by a hypothesis based on the structure of the MUP genes. We suggest that the protein agent is MUP, the active part of the molecule being the six N-terminal amino acids, and that the dialysable agent is the hexapeptide coded for by the Group 2 genes. Proteolysis of the protein agent would release dialysable fragments containing the N-terminal hexapeptide. The sequences of the two hexapeptides are quite similar: Group 1, N-Glu-Glu-Ala-Ser-Ser-Thr; Group 2, N-Glu-Glu-Ala-Arg-Ser-Met.

Group 1 and Group 2 genes have randomly diverged

BS6 and BS2,3 are members of the two major groups of MUP genes in the BALB/c genome. The numbers of Group 1 and Group 2 genes are approximately equal (Bishop *et al.*, 1982). This is because the predominant organisation of the MUP locus is an array of 45 kb domains each containing a Group 1 and a Group 2 gene linked in a divergent orientation (Clark *et al.*, 1984b; Bishop *et al.*, 1985). We have presented the sequence of BS6 and BS2,3 over a homologous region ~4.5 kb in length that includes the entire transcription unit as well as 5′ and 3′ flanking sequences. The most obvious differences between the two sequences are the three long insertions/deletions. In each case these occur in regions of 'simple sequence' DNA suggesting that they may have been created by 'slippage' during DNA synthesis or repair (Ghosal and Saedler, 1978). In general, the divergence between the two sequences is uniformly spread across the region sequenced (Table IV). Thus no recent gene correction has occurred between the two genes such as has been observed between human $^{G}\gamma$ and $^{A}\gamma$ globin genes (Slightom *et al.*, 1980). The exons and introns of BS6 and BS2,3 have diverged to about the same extent. Comparisons between other active genes indicate that, in general, intronic sequences diverge more rapidly than exonic sequences (Perler *et al.*, 1980; Efstratiadis *et al.*, 1980) presumably because introns have lesser selective constraints acting on them. That this is not the case in the comparison of the two MUP genes possibly indicates that the ancestral BS2,3 pseudogene was free to diverge at the same rate in both introns and exons. Group 2 genes, however, are reasonably well conserved amongst themselves and we have drawn from this observation the conclusion that the 45-kb domain, rather than the individual MUP gene, is the unit of evolutionary change of the majority of MUP genes (Clark *et al.*, 1984b; Bishop *et al.*, 1985).

Table IV. Divergence between BS6 and BS2,3

Region	Divergence (%)
Full length	13.4
5′ flanking region	11.1
Transcription unit	12.6
mRNA	13.1
Translated mRNA	11.5
Non-translated mRNA	15.5
Intronic sequences	13.1
3′ flanking region	15.6

The divergence between the two genes was estimated over the regions indicated. In this analysis each base change was scored as 1, as was each insertion/deletion, irrespective of size.

Materials and methods

Cloned DNA

The isolation of MUP genomic clones and subclones is described in Clark *et al.* (1982, 1984b) and Bishop *et al.* (1982). The propagation of bacteriophage and plasmid clones and the isolation of DNA were carried out as described (Clissold and Bishop, 1982; Clark *et al.*, 1982; Bishop *et al.*, 1982).

DNA sequencing

To obtain the complete 4 kb sequences of BS6 and BS2,3, fragments of plasmid pBS6-2, pBS6-5, pBS6-1-1, pBS2-2 and pBS3B-3 were cloned into M13mp7, 8 or 9 and sequenced by the dideoxy nucleotide method, essentially as described by Sanger *et al.* (1977) and Anderson *et al.* (1980). Two main strategies were employed to ensure that continuous stretches of sequence would be generated. (i) The cloned fragments were digested with restriction enzymes that cleave 4 bp recognition sites and 'shotgunned' into M13 vectors. (ii) Larger subfragments were cloned into M13mp8, replicative forms were prepared and a second generation of M13 clones containing progressively shorter fragments was isolated by the method of Hong (1982).

S1 nuclease protection

The probe was a 696 bp *Alu*I fragment, extending from nucleotide +127 to nucleotide −568 in the BS6 sequence (Figure 3), and cloned at the *Hinc*II site of M13mp7. The single-stranded M13 clone was annealed to the sequencing primer and the strand complementary to MUP mRNA was uniformly labelled using the Klenow fragment of DNA polymerase I (Boehringer). The double-stranded region thus created was digested with *Eco*RI and the fragment lying between the two *Eco*RI cloning sites of the vector (sp. act. 10^7-10^8 c.p.m./μg) was purified on a 5% polyacrylamide gel. An aliquot of the probe (20 000 c.p.m.) was co-precipitated with 1 μg of total poly(A)$^+$ RNA and redissolved in 10 μl of 40 mM Pipes (pH 6.4), 1 mM EDTA, 0.4 M NaCl, 80% formamide. Samples were denatured at 85°C for 15 min and incubated at 50°C for 4 h. Samples were digested with 250 U/ml S1 nuclease (Sigma) at 37°C for 1 h in 100 μl of 0.28 M NaCl, 0.05 M NaAc (pH 4.6), 4.5 mM $ZnCl_2$ and 10 μg/ml single-stranded salmon sperm DNA, phenol extracted, precipitated twice with ethanol and resuspended in 3 μl of formamide dye mix.

Primer extension (from Ghosh et al., 1981)

The primer extension probe was the 93 bp *Alu*I-*Sau*961 fragment between nucleotides +34 and +127 in the BS6 sequence (Figure 3). This was prepared and annealed to poly(A)$^+$ RNA in essentially the same way as the S1 protection probe (above). Annealing was terminated by the addition of 250 μl ice-cold

0.3 M NaAc (pH 7.0) followed by two ethanol precipitations. The pellet was resuspended in 50 μl 50 mM Tris-HCl (pH 8.3), 6 mM $MgCl_2$, 40 mM HCl, 10 mM DTT with 1 mM of each deoxynucleotide triphosphate and 1 unit of AMV reverse transcriptase was added. After equilibration on ice for 5–10 min, the reaction mixture was incubated for 3 h at 37°C. NaOH was then added to 0.2 N and the incubation continued for a further 1 h. The reaction mixture was neutralised with 10 N HCl, phenol extracted, and ethanol precipitated twice. Pellets were resuspended in 3 μl of formamide dye mix and loaded on 6% sequencing gels.

Acknowledgements

We thank Morag Robertson and Melville Richardson for technical assistance, and the MRC and Cancer Research Campaign for financial support.

References

Anderson,S., Gait,M.J., Mayol,L. and Young,I. (1980) *Nucleic Acids Res.*, **8**, 1731-1743.

Bishop,J.O., Clark,A.J. Clissold,P.M., Hainey,S. and Francke,U. (1982) *EMBO J.*, **1**, 615-620.

Bishop,J.O., Selman,G.G., Hickman,J., Black,L., Saunders,R.D.P. and Clark, A.J. (1985) *Mol. Cell. Biol.*, **5**, 1591-1600.

Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.

Clark,A.J., Clissold,P.M. and Bishop,J.O. (1982) *Gene*, **18**, 221-230.

Clark,A.J., Clissold,P.M., Al-Shawi,R., Beattie,P. and Bishop,J.O. (1984a) *EMBO J.*, **3**, 1045-1052.

Clark,A.J., Hickman,J. and Bishop,J.O. (1984b) *EMBO J.*, **3**, 2055-2064.

Clark,A.J., Chave-Cox,A., Ma,X. and Bishop,J.O. (1985) *EMBO J.*, **4**, 3167-3171.

Clissold,P.M. and Bishop,J.O. (1982) *Gene*, **18**, 211-220.

Crouse,G.F., Simonsen,C.C., McEwan,R.N. and Schimke,R.T. (1982) *J. Biol. Chem.*, **256**, 8407-8415.

Devereux,J. and Haeberli,P. (1984) Program Library of the University of Wisconsin Genetics Computer Group.

Dickson,L.A., Ninomya,Y., Bernard,M.P., Pesciotta,D.M., Parsons,J., Green,G., Eikenberry,E.F., de Crombrugghe,B., Vogeli,G., Pastan,I., Fietzek,P.P. and Olsen,B.R. (1981) *J. Biol. Chem.*, **256**, 8407-8415.

Efstratiadis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritz, R.A., De Rich,J.K., Forget,G.G., Weissmann,S.M., Slightom,J.L., Blechl, A.E., Smithies,O., Baralle,F.E., Shoulders,C.S. and Proudfoot,N.J. (1980) *Cell*, **21**, 653-668.

Eiferman,F.E., Young,P.R., Scott,R.W. and Tilgham,S.M. (1981) *Nature*, **294**, 713-718.

Evans,G.A., Margulies,D.H., Camerini-Otero,R.D. and Seidman,J.G. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 1994-1998.

Gerin,M.B., Cooper,D.L., Eiferman,F., Van de Rijn,P. and Tilghman,S.M. (1981) *J. Biol. Chem.*, **256**, 1954-1959.

Ghazal,P., Clark,A.J. and Bishop,J.O. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4182-4185.

Ghosal,D. and Saedler,H. (1978) *Nature*, **275**, 611-617.

Ghosh,T.K. Reddy,V.D., Taitak,M., Lebowitz,P. and Weissmann,S.M. (1981) *Methods Enzymol.*, **65**, 580-594.

Glanville,N., Durnam,D.M. and Palmiter,R.D. (1981) *Nature*, **292**, 267-269.

Hagenbuechle,O., Tosi,M., Schibler,U., Bovey,R., Wellauer,P.K. and Young, R.A. (1981) *Nature*, **289**, 643-646.

Hong,G.F. (1982) *J. Mol. Biol.*, **158**, 539-549.

Keller,E.B. and Noon,W.A. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 7417-7420.

Konkel,D.A., Maizel,J.V. and Leder,P. (1979) *Cell*, **18**, 865-873.

Kozak,M. (1984a) *Nucleic Acids Res.*, **12**, 857-872.

Kozak,M. (1984b) *Nature*, **308**, 241-246.

Kramer,A., Keller,W., Appel,B. and Luehrmann,R. (1984) *Cell*, **38**, 299-307.

Kuhn,N.J., Woodworth-Gutai,M., Gross,K.W. and Held,W.A. (1984) *Nucleic Acids Res.*, **12**, 6073-6090.

Law,S.W. and Dugaiczyk,A. (1981) *Nature*, **291**, 202-205.

Lerner,M.R., Boyle,J.A., Mount,S.M., Wolin,S.L. and Steitz,J.A. (1980) *Nature*, **283**, 220-224.

Mishioka,Y. and Leder,P. (1979) *Cell*, **18**, 875-882.

Moore,K.W., Sher,B.T., Sun,Y.H., Eakle,K.A. and Hood,L. (1982) *Science (Wash.)*, **215**, 679-682.

Mount,S.M., Pettersson,I., Hinterberger,M., Karmas,A. and Steitz,J.A. (1983) *Cell*, **33**, 509-518.

Nunberg,J.H., Kaufman,R.J., Chang,A.C.Y., Cohen,S.N. and Schimke,R.T. (1980) *Cell*, **19**, 355-364.

Perler,F., Efstratiadis,A., Lomedico,P., Gilbert,W., Kolodner,R. and Dodgson, J. (1980) *Cell*, **20**, 555-565.

Rogers,J. and Wall,R. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 1877-1879.

Ruskin,B., Krainer,A.R., Maniatis,T. and Green,M.R. (1984) *Cell*, **37**, 415-427.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.

Simonsen,C.C. and Levinson,A.D. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 2495-2499.

Slightom,J.L., Blechl,A.E. and Smithies,O. (1980) *Cell*, **21**, 627-638.

Vandenbergh,J.G., Whitsett,J.M. and Lombardi,J.R. (1975) *J. Reprod. Fertil.*, **43**, 515-523.

Vandenbergh,J.G., Finlayson,J.S., Dobrogosz,W.J., Dills,S.S. and Kost,T.A. (1976) *Biol. Reprod.*, **15**, 260-265.

Weiss,E., Golden,L., Zakut,R., Mellor,A., Fahrner,K., Kvist,S. and Flavell,R.A. (1983) *EMBO J.*, **2**, 453-462.

Wozney,J., Hanahan,D., Boedtker,H. and Doty,P. (1981) *Nature* **294**, 129-135.

Young,R.A., Hagenbuechle,O. and Schibler,U. (1981) *Cell*, **23**, 451-458.

Received on 27 June 1985; revised on 16 September 1985