RAT PROSTATIC STEROID BINDING PROTEIN:

ORGANIZATION AND EXPRESSION OF THE TWO C3 GENES

Helen Catherine Hurst

A thesis submitted for the degree of
Doctor of Philosophy

Imperial College of Science and Technology, London

To Matthew and my parents

ABSTRACT

Rat prostatic steroid binding protein, whose expression is stimulated by androgens, consists of two subunits. One contains the polypeptides C1 and C3 while the other contains the polypeptides C2 and C3. Two non-allelic genomic clones specific for C3 have been isolated from rat DNA libraries constructed in bacteriophage λ. These clones, termed C3(1) and C3(2), have been extensively mapped using restriction enzymes and R-loop analysis and subcloned into M13 vectors for sequencing by the Sanger chain termination method. Both genes are ~3.0 kb, each with two introns at identical positions and their coding regions share 97.7% sequence homology. However, only C3(1) is totally compatible with the published C3 protein sequence. This observation, coupled with direct studies on in vivo C3 transcription, suggests that C3(1) alone is transcribed in the ventral prostate while C3(2) is probably silent.

Transcript mapping has shown that there are two distinct C3 transcripts which share a unique 3' terminus, but have 5' termini 38 bases apart, each preceded by a 'TATA' box homology. The two transcripts are produced at a steady state ratio of 18:1 in normal and hormonally manipulated rats.

Upstream of these C3 cap sites a short Alu-like repetitive element has been found, between -450 and -370 in both C3 genes. This sequence shows strong homology to other isolates of rat Alu elements and together they form a new third class of rodent Alu-equivalent sequences which probably fulfil no useful function in the genome and are examples of selfish DNA.

It is hoped that this characterization of the C3 genes will provide the basis of further experiments to investigate the mechanism of androgen controlled gene expression using C3 as a 'model'.

ACKNOWLEDGEMENTS

I should like to thank Roger White, Maurice Needham and Martin Page for being so tolerant as to work in the same laboratory as me for three years - you can now all leave nasty things on my bench without fear of retribution! Many other people at the I.C.R.F. have been of great assistance to me during my bursarship and I particularly acknowledge the kind interest and help of Shelagh Redmond, Mick Jones, Gordon Peters, Clive Dickson, John Arrand, Colin Thacker and Trudi Veldmann,

Thanks are also due to David Glover at Imperial College for making the bureaucracy almost invisible, while Jackie Needham deserves a medal for being able to decipher my writing and type it all so nicely. I am grateful to Audrey Symons and the I.C.R.F. Photographic Department for producing the figures and being so patient when I kept changing my mind!

However, the biggest thanks of all must go to Malcolm Parker for supervising me through all my 'megadepressions' and over-excitements without apparently losing patience.

H.C.H.

ABBREVIATIONS

b	bases
bp	base pairs
kbp	kilobase pairs
ss	single stranded
ds	double stranded
cDNA	complementary or copy DNA
dNTP	2' deoxyribonucleotide triphosphates
ddNTP	2', 3' dideoxyribonucleotide triphosphates
Mr	Molecular ratio
DW	Distilled Water
o/n	overnight
mins	minutes
RT	room temperature
U.V.	ultraviolet light
r.p.m.	revolutions per minute
c.p.m.	counts per minute
w/v	weight to volume
phage	bacteriophage
p.f.u.	plaque forming units
OAc	acetate
PPi	pyrophosphate
DTT	Dithiothreitol
SDS	Sodium dodecyl sulphate
MMTV	Mouse mammary tumour virus
LTR	Long terminal repeat
S-Rc	Steroid-receptor complex
RE	Repetitive element
SV40	Simian virus 40
tk	thymidine kinase gene.

CONTENTS

	Page
ABSTRACT	3
ACKNOWLEDGEMENTS	4
ABBREVIATIONS	5
SECTION I: INTRODUCTION	10-28
1. Theoretical model for steroid action.	10
2. Features of gene structure and expression in eucaryotes.	12
3. Molecular approaches to the study of steroid hormone action.	15
4. The rat ventral prostate as a model for the study of androgen action.	22
5. Interspersed repetitive DNA.	24
SECTION R: RESULTS	29-103
R1. Isolation and characterization of C3 genomic clones.	30
R2. DNA sequence analysis of C3(1) and C3(2).	46
R3. Mapping of C3 transcripts to C3(1).	63
R4. Is C3(2) expressed in vivo?	78
R5. Characterization of the C3 Alu-like repetitive element.	88
SECTION D: DISCUSSION	104-122
1. Structure of the C3 genes.	104
2. C3 expression in vivo.	106
3. The Alu-like repetitive element.	111
4. Control of C3(1) gene expression.	114

SECTION M: MATERIALS AND METHODS 123-146

M1. Materials 123

M2. Methods 130

REFERENCES 147-154

APPENDIX: Details of genomic subclones. 155-162

LIST OF TABLES

		Page
1.	A survey of chimeric genes constructed to study the action of steroid hormones.	18
2.	Survey of inducible genes whose required sequences have been mapped.	19
3.	Homologies to the chicken progesterone receptor-DNA binding site.	119
4.	Contents of 'dideoxysequencing' nucleotide mixes.	143
5.	Primer extension sequencing mixes.	145

LIST OF FIGURES

1.	Proposed scheme to explain the apparent mobility of Alu elements.	27
2.	Analysis of 13 isolates from λ Charon 4A rat DNA libraries.	33
3.	Restriction enzyme mapping of λ5D.	33
4.	Comparison of C3 sequences isolated as λ6 and λ8 with those present in rat cell DNA.	36
5.	Comparative restriction enzyme digests of p61 and p81.	38
6.	Restriction enzyme maps of the two C3 genes.	40
7.	Analysis of R-loops in the electron microscope.	43
8.	DNA sequencing strategy.	51
9.	Maxam and Gilbert sequence analysis.	54
10.	DNA sequence of two C3 genes.	56
11.	Comparison of intervening sequences from the two C3 genes.	58
12.	Second exon differences between C3(1) and C3(2).	61
13.	Mapping 5' ends of prostatic C3 transcripts.	65
14.	Primer extension sequence analysis of C3 transcripts.	68
15.	Mapping 3' ends of prostatic C3 transcripts.	71
16.	Densiometric scans of S1 and primer extension 5' mapping gel autoradiographs.	73

Figures (cont'd.)

17. Effect of androgens on C3 transcript levels. 75

18. Analysis of the five available C3 cDNA clones. 80

19. *In vitro* transcription of the C3 genes. 83

20. Direct probing for *in vivo* C3(2) mRNA. 85

21. A comparison of five short, Alu-like rat repetitive elements. 90

22. Assaying for Alu transcripts. 94

23. Mapping 5' and 3' termini of Alu transcripts. 96

24. Features of the Alu element within the C3 genes. 99

25. Further analyses of transcription from the Alu element. 101

26. Summary of the state of DNA methylation within and flanking the two C3 genes. 110

27. Features within the proximal 5' flanking DNA of the C3(1) gene. 117

28. Subclones of the C3 genes. 157

29. Generation of C3(1) 5' deletion mutants. 160

SECTION I: INTRODUCTION

Virtually all tissues of mammalian organisms respond to at least one steroid hormone. Many of these responses are extremely specific, resulting in the elevated synthesis of a limited set of proteins within a defined target tissue. Steroids also elicit other responses of a general or 'pleiotypic' nature (Hershko et al., 1971) involved in the regulation of growth.

This thesis sets out to establish a working 'model' to study the mechanism of action of the specific response to steroids, which itself is a special case of the more general problem as to how gene expression is controlled during cellular differentiation and development.

1. Theoretical Model for Steroid Action

A general model to explain how steroids exert a specific effect on their target tissues has been expanded from ideas first proposed to account for the interaction of oestrodiol within rat uterine cells (Jensen et al., 1968).

i) Steroid enters cells by passive diffusion.

ii) Steroid binds to specific receptors present only in the cytosol of target tissues.

iii) The hormone-receptor complex (S-Rc) translocates to the nucleus.

iv) An ill-defined activation process occurs which is believed to result in direct binding of the complex to a finite number of acceptor sites on the chromatin, so directly increasing transcription from a specific set of genes whose protein products thereafter accumulate within the target tissue.

After fifteen years there is, strictly speaking, no direct proof of this model, but many lines of evidence have accumulated to support each proposed step as discussed below.

The central dogma of Jensen's model is that steroids accumulate within their target tissues solely because these cells possess the correct receptors and it is the S-Rc itself that elicits the steroid response. These ideas are supported by the lack of evidence for specific steroid transport mechanisms within target cell membranes

(reviewed Gorski and Gannon, 1976; Higgins and Gehring, 1978). Thus, steroids can diffuse equally into all cells, through their lipophillic cell membranes. Also, the steroid itself seems to be the only signal received by responsive cells. A requirement for a second messenger, such as cyclic AMP, has been examined, but discounted (reviewed Higgins and Gehring, 1978). Moreover, the biological activity of steroids does not generally depend on their metabolic transformation to active inducers within target cells (Jensen and Jacobsen, 1962). The only exception to this is for the androgen, testosterone, which is converted by a specific reductase, within male sex accessory tissues to an active, receptor-binding form (Wilson and Gloyna, 1970).

The rat uterine oestrogen receptor was the first receptor protein isolated in both cytosolic (Toft and Gorski, 1966) and nuclear (Jensen et al., 1967) forms. Since then the properties and distribution of a variety of receptor molecules have been studied using radiolabelled steroid and, more recently, monoclonal antibodies. In general, the characteristics determined for receptor molecules are compatible with the role proposed for them. For example, physiological levels of steroids within the bloodstream are quite low (10^{-10} to 10^{-7}M, depending on the steroid class) and receptor levels within cells show similar concentrations. Also, steroids bind a variety of proteins in a non-specific manner (Kd = 10^{-5} to 10^{-4}M), but binding to receptor is very tight and specific (Kd = 10^{-10} to 10^{-9}M) (reviewed Higgins and Gehring, 1978). This receptor-steroid binding probably constitutes a simple reverse equilibrium: Rc+S = S-Rc, but could be more complicated as receptors may all prove to be oligomeric proteins, making a co-operative binding mechanism possible. To date only the chicken progesterone receptor has been shown to possess two subunits. These are similar, but distinct polypeptides, being the products of separate genes (Schrader et al., 1981).

How the S-Rc complexes enter the nucleus is still a matter of debate, but it has long been thought that once there they bring about a direct increase in gene transcription. One of the first effects of oestrogen on the rat uterus was shown to be a stimulation of nuclear RNA synthesis (Hamilton et al., 1965) which could be blocked by actinomycin, but was insensitive to cycloheximide (Means and Hamilton, 1966). A more 'visual' demonstration of the transcriptional response to steroids was provided by the action of ecdysone, which is released at the end of larval development in Drosophila. When salivary gland

explants were exposed to ecdysone a specific pattern of chromosomal puffs was observed (Ashburner et al., 1973). This puffing pattern, associated with localized high rates of transcription, was rapidly established (within five minutes of hormone addition) and was insensitive to a variety of inhibitors of protein synthesis. Compelling evidence that S-Rc complexes were interacting with these nuclear transcription units was provided by Gronemeyer and Pongs (1980). Using a 'high affinity' ecdysone analogue, the steroid was photocross-linked in situ to the polytene chromosomes. Covalently bound hormone was then located by an indirect immunofluorescence microscopy assay and found to be bound at the sites of ecdysone-induced chromosomal puffs.

Therefore it appears that specific nuclear acceptor sites for S-Rc complex binding are adjacent to the particular genes whose products accumulate during a response to steroid. To be able to elucidate the mechanism of this gene activation it is necessary to understand something of eucaryotic gene structure and expression.

2. Features of Gene Structure and Expression in Eucaryotes.

The exon/intron structure found for the majority of eucaryotic protein-encoding genes is now a well-known concept. These genes are transcribed by RNA polymerase II, which is sensitive to low levels of α-amanitin (Weinmann et al., 1974). Transcription initiation occurs at a defined point on a gene termed nucleotide +1, though for innumerable viral and cellular genes it has now been established that several different, but specific initiations may occur within a few bases, producing a family of transcripts (reviewed Nevins, 1983). In these cases the initiation nucleotide of the most abundant mRNA is termed +1. Soon after initiation a m^7Gp 5' 'cap' is added to the first nucleotide of the growing chain (reviewed Darnell, 1982). Termination of transcription is poorly understood, but certainly proceeds beyond the polyadenylation site, as no case of definite termination here has been found in studies using pulse-labelled RNA (reviewed Darnell, 1982). Therefore the 3' polyA tail is added post-transcriptionally after an endonucleolytic cleavage some 10-25 bases downstream of a putative signal AAUAAA (Proudfoot and Brownlee, 1976). Other sequences around this hexanucleotide have also been thought important (Benoist et al., 1980), but only the AAUAAA sequence so far has been shown to be functionally recognised (Fitzgerald and Shenk, 1981).

Splicing out of the intervening sequences present in the primary transcript usually, but not necessarily, occurs after polyadenylation (reviewed Darnell, 1982). The precise mechanism by which splicing occurs is still a matter of conjecture, but the importance of sequences flanking the intron/exon boundaries of most genes (Breathnach et al., 1978; Mount, 1982) has been functionally demonstrated in several ways. Perhaps the most 'natural' illustration has been the discovery that the occurrence of various β-thalassemias may be ascribed to point mutations at the splice sites within the β-globin genes of particular patients (reviewed Spritz and Forget, 1983).

Mechanisms by which the binding of RNA polymerase is directed to the beginning of a gene and further processes controlling the rate of transcriptional initiation have also yet to be determined. By analogy with procaryotic systems, these elements of a gene, termed 'promoters', are thought to reside chiefly within the 5' flanking DNA. These sequences are numbered away from the initiation nucleotide at +1 such that the base immediately 5' of this is termed -1 and the negative number increases with distance 5' from the cap site.

The first putative eucaryotic promoter sequence identified was the 'TATA'-box, an AT-rich region centred at -25 to -30 in many genes (reviewed Breathnach and Chambon, 1981). This sequence is thought to be functionally analogous to the polymerase binding site or Pribnow box (5' TATAATG 3') found at -10 in procaryotic genes (reviewed Rosenberg and Court, 1979). Another sequence believed to be important for procaryotic transcription is found further upstream around -35. This is termed the recognition region and probably plays a role in transcriptional modulation (Siebenlist et al., 1980; reviewed Rosenberg and Court). Eucaryotic genes have also been examined for upstream elements and a consensus sequence 5' GG$^{C}_{T}$CAATCT 3' centred at -70 to -80, the so-called 'CAAT'-box, has been identified in a number of genes (Benoist et al., 1980; reviewed Breathnach and Chambon, 1981), but is not as ubiquitous as the 'TATA'-box.

Several systems for analysing the transcription of cloned genes, away from their in vivo environment, have now been developed. These fall into two groups: cell free systems and systems whereby cloned genes are transfected into suitable cultured cells. Both of these assays have been used to assess the importance of 5' flanking sequences for efficient and accurate transcription from a great many genes. This has lead to the conclusion that, certainly in cell free

in vitro systems, the 'TATA'-box is essential for efficient initiation at the *in vivo* cap sites (e.g. for rabbit β-globin, Grosveld *et al.*, 1981). Deletion of the sequence can lead to a loss of detectable transcription (as for the rabbit β-globin gene), or often a slight reduction in efficiency coupled with a total loss of accuracy with multiple, heterogeneous starts occurring (e.g. for SV40 early genes, Mathis and Chambon, 1981). In whole cell assays the 'TATA'-box is not generally absolutely required. Its loss may also result in heterogeneous starts not seen *in vivo* (Grosveld *et al.*, 1982; Benoist and Chambon, 1981), but in some experiments no loss of accuracy or efficiency is observed (Dierks *et al.*, 1983) and it is sequences further upstream that are essential.

The importance of upstream sequences has now been functionally demonstrated, by assessing the transcriptional capability of deletion mutants (made *in vitro*) in expression systems, for a number of viral and cellular genes. In general there appear to be one or two such regions 40-100bp further upstream of the 'TATA'-box. They often contain (imperfect) inverted repeats or dyad symmetrics as found for the *Drosophila* heat shock genes (Pelham, 1982), the Herpes thymidine kinase gene (McKnight, 1982), the mouse metallothionein -1 gene (Brinster *et al.*, 1982) plus several others. In some genes direct repeats seem to be important, as for the -100 region of rabbit β-globin (Dierks *et al.*, 1983) and the 21bp repeats of SV40 (Everett *et al.*, 1983). Many of these repeats contain GC-rich sequences.

Certain viruses have been found to contain additional upstream sequences which appear to increase the transcriptional efficiency of almost any attached gene, independently of their orientation and position with respect to that gene (reviewed Khoury and Gruss, 1983). These elements have been termed enhancers and sequences with similar properties have also been found in cellular immunoglobulin genes (reviewed Boss, 1983).

The progess made in understanding eucaryotic gene transcription was largely due to advances in molecular cloning and DNA sequencing. Such molecular approaches are also being applied to the particular case of steroid regulation of gene transcription.

3. Molecular Approaches to the Study of Steroid Hormone Action.

The availability of specific cloned probes made it possible to directly assay in vivo levels of steroid inducible transcripts to assess the effect and likely site of action of the hormone. Certainly, in all the systems currently under investigation, a stimulation in specific mRNA levels in response to steroid can be directly measured. For example, the synthetic glucocorticoid, dexamethasone, induced within 15 minutes a 10-fold increase in levels of mouse mammary tumour virus (MMTV) RNA in mouse mammary carcinoma cells (Ringold et al., 1977; Young et al., 1977). The rapidity of this response strongly suggests a direct effect of steroid at the level of gene transcription.

Another glucocorticoid induced gene, mouse metallothionein 1 (MT-1),also responds rapidly, a 90-fold increase in transcription being recorded within 2 hours in isolated mouse liver cells. This increase was sensitive to low levels of α-amanitin (Hager and Palmiter, 1981). However, results obtained from a variety of mouse cell lines suggested that dexamethasone had a 5-fold effect on MT transcription plus a 2-fold effect on MT-RNA stability (Mayo and Palmiter, 1981).

In other systems too, the increase in mRNA levels in response to steroid cannot usually be explained by steroids acting solely as transcriptional inducers. For example, the tubular gland cells of the mature chicken oviduct will respond to a variety of steroids, producing increased levels of transcription from the egg white protein genes, ovalbumin, conalbumin, ovomucoid and lysozyme (reviewed Higgins and Gehring, 1978). By examining mRNA levels in isolated oviduct nuclei from control and steroid-treated chicks, it was concluded that while levels of newly synthesized ovalbumin RNA ($mRNA_{ov}$) rose due to increased transcription (Swaneck et al., 1979) there was also an effect on mRNA stability. In the absence of steroid the half-life of $mRNA_{ov}$ declined from >20 hrs to 4.5 hrs (Palmiter and Carey, 1974), while conalbumin RNA half-life was shown to decline from 8 hrs to 3 hrs (McKnight and Palmiter, 1979).

Therefore, two questions still remain: i) are there specific DNA sequences or structures around inducible genes that render them sensitive to hormones,and ii) do these specific sequences show high affinity for S-Rc complex binding leading to transcriptional activation of the gene?

To help answer these questions various cloned inducible genes have now been transfected into receptor-containing heterologous cells. Generally some measure of specific, steroid-inducible transcription was retained (reviewed Coffino, 1981). This implied that the ability to respond to steroid transferred with the cloned DNA, but did not rule out the possibility that steroids interacted with gene transcripts, increasing their stability and therefore their abundance. Moreover, such studies were limited by the fact that few cultured cell lines retain steroid receptors. It is therefore desirable to develop a way of introducing specific genes into primary cultures of homologous cells. For this reason a variety of chimeric genes have been constructed where the 5' end (presumably containing the regulatory regions) was provided by the inducible gene, but the bulk of the expressed sequences were from a non-responsive gene. Examples of such genes are given in Table 1.

The majority of these fused genes were constructed with the 'join' within the sequences encoding 5' untranslated mRNA. Therefore the resultant chimeric transcripts will bear 5' sequences originating from the responsive gene. For this reason some of the experiments may be criticised as the observed inducibility of the fused mRNA's could be due to an interaction between the S-Rc complex and these 5' sequences. However, in some experiments the two genes were joined either very close to the cap site or within their promoter regions (Table 1, nos. 3,6,7). The continued observation of hormone-inducible transcripts in these cases really implies that the 5' flanking DNA sequences of inducible genes can confer the ability to transcriptionally respond to steroids.

This conclusion was supported by further studies, where the inducible region of flanking DNA was delineated using deletion mutants of the chimeric genes. A summary of the findings to date is presented in Table 2. It should be noted that the sequences believed important for the glucocorticoid induction of rat α_{2u} globulin were deduced by comparing the inducible gene sequence with that of constitutive variants. These variant genes were expressed at high levels (similar to the induced level) in the absence of hormone and were found to have 11bp (from -88 to -77) missing from the 5' flanking DNA sequence. It is therefore interesting that within the inducible gene these nucleotides form part of a 9bp element directly repeated further downstream, as indicated in Table 2. Buetti and Digglemann (1983) also noticed

Footnotes to Table 1:

MMTV = mouse mammary tumour virus; HSV = herpes simplex virus; tk = thymidine kinase; DHFR = dihydrofolate reductase; HaMuSV = Harvey murine sarcoma virus; p21 = transforming protein of HaMuSV; Dex = dexamethasone; prog = progesterone; oest = oestrogen; CHO = Chinese hamster ovary cells.

Table 1. A survey of chimeric genes constructed to study the action of steroid hormones.

	Inducible gene	Amount of 5' flanking DNA	Non-inducible gene	Inducibility of transfected chimeric genes	Reference
1	MMTV	-1000 to +270	Mouse DHFR	3-5 fold with dex in CHO *dhfr*$^-$ cells	Lee *et al.*, 1981
2	MMTV	-1000 to +105	HaMuSV p21 gene	10-50 fold with dex in NIH 3T3 cells	Huang *et al.*, 1981
3	Human growth hormone	-500 to +3	HSV tk	3-5 fold with dex in mouse *Ltk*$^-$ cells	Robins *et al.*, 1982
4	Chicken lysozyme	-1400 to +15	SV40 T-antigen	5-10 fold with dex or prog, but not with oest. in primary cultures of chicken oviduct cells	Renkawitz *et al.*, 1982
5	Rat α_{2u} globulin	-235 to +68	HSV tk	2-4 fold with dex in mouse *Ltk*$^-$ cells	Kurtz *et al.*, 1982
6	MMTV	-1000 to +435	HSV tk (from -77)	Two inducible transcripts observed: 1 chimeric, the other tk mRNA. Both respond 5 fold with dex in mouse Ltk$^-$ cells	Hynes *et al.*, 1983
7	MMTV	-449 to -109	HSV tk (from -109)	3-16 fold with dex; 1 case of 50 fold. Tk mRNA only produced. Rat XC*tk*$^-$ cells	Chandler *et al.*, 1983
8	MMTV	-600 to +134	HSV tk	10-20 fold with dex in mouse Ltk$^-$ cells	Buetti and Digglemann, 1983
9	Chicken ovalbumin	-753 to +41	Chicken β-globin	10 fold with prog in primary cultures of chicken oviduct cells	Dean *et al.*, 1983

Table 2. Survey of inducible genes whose required sequences have been mapped by observation of the response to steroid by a series of 5' deletion mutants.

	Inducible gene	Hormone(s)	5' sequences required for inducibility	Reference
1.	MMTV	Glucocorticoid	to -202	Hynes et al., 1983
2.	MMTV	Glucocorticoid	-204 to -105 5' TGGTTT 3'	Buetti and Digglemann, 1983
3.	Chicken ovalbumin	Progesterone	-95 to -222	Dean et al., 1983
4.	Chicken lysozyme	Glucocorticoid Progesterone	-140 to -200	Renkawitz et al., 1983
5.	Rat α2u globulin	Glucocorticoid	-91 to -83 -68 to -60 5' TGAGAAG$^{A}_{G}$A 3'	Kurtz et al., 1982

the occurrence of direct repeats within the required region of the MMTV long terminal repeat (LTR). Furthermore, they found that the smallest 5' deletion they examined (to -105) retained transcription levels reminiscent of the non-induced, intact gene. They therefore concluded that glucocorticoid responsive sequences could be physically and functionally separated from 'promoter' elements ('TATA'-boxes?) that merely direct correct initiation of transcription. This conclusion was also reached by Chandler et al. (1983) who additionally provided evidence that the MMTV hormone inducible sequences might act as a hormone-activated enhancer element, functioning in either orientation. Such an observation has not been repeated by other groups though (Dr. G. Ringold, personal communication) and anyway seems at variance with results from α_{2u} globulin. These implied that hormone responsive sequences are actual repressors of transcription and that steroids act by relieving this repression.

Deletion mapping studies have partly delineated hormone inducible sequences within the first ~250bp of 5' flanking DNA of several responsive genes. They do not, however, prove a direct interaction between the S-Rc complex and the steroid-responsive sequences. Therefore, to examine this issue, several groups have looked at the in vitro binding of S-Rc complexes to cloned DNA. Currently two assays are used: one is a competition assay with cellulose-bound calf thymus DNA; the other is a nitrocellulose binding assay (reviewed Cato, 1983). Neither technique produced unequivocal results, but use of both assays has achieved some consensus of opinion.

Certainly for MMTV, several groups have located glucocorticoid binding sites within the LTR using small DNA restriction fragments and both S-Rc binding assays (Pfahl, 1982; Geisse et al., 1982), in one case coupled with electron microscopy (Govindan et al., 1982). There seem to be two or three binding regions located in the 300bp between -100 and -400 (Payvar et al., 1982). More exact localization of the binding sites has been attempted by determining the sequences protected from DNase 1 by S-Rc binding (Scheidereit et al., 1983). Two protected regions were found (from -72 to -124 and from -163 to -192) which share some sequence homology, particularly the hexanucleotide 5' TGTTCT 3'. Therefore these results generally support the findings from deletion mutant analysis which also concluded that up to 200bp of flanking DNA were required for inducibility in MMTV (see Table 2).

Similar experiments have been performed in the study of the interaction of the chicken progesterone receptor with the egg-white protein genes, but there is less agreement between different groups. Using the competition binding assay, Mulvihill et al. (1982) found specific binding to a region 250 to 300bp upstream of the ovalbumin gene cap site. Other fragments of DNA from the ovomucoid, conalbumin and ovalbumin related X and Y genes were also positive in the assay. The sequences of these fragments were compared and a consensus sequence derived: 5' ATC$^{CC}_{TT}$ATT$^{A}_{T}$TCTG$^{G}_{T}$TTGTA 3'. This sequence was found repeated several times within the 5' flanking DNA (between -100 and -1000) of all five genes, though conalbumin had only one copy. In contrast, Compton et al. (1983) using the filter binding assay, concluded that preferential binding occurred closer to the ovalbumin cap site, between -135 and -247. This finding was strengthened by nuclease protection experiments showing the S-Rc complexes covered a region from -155 to -195 (Schrader et al., 1983). More recent studies from Chambon's group now lead them to conclude that the progesterone-receptor complex can bind at three distinct places on the conalbumin gene between +62 and -102. They further believe that repeated sequence elements, 5' GCTXY 3' (where X and Y are usually C) are important for this binding (Davison et al., 1982).

Clearly then, recent studies on the 5' flanking DNA of steroid-responsive genes have highlighted several regions and sequences that may play a functional role. Obviously these sequences must be assayed before their full significance can be assessed. However, at this stage one may tentatively propose that a pattern of very short repeated sequences and multiple S-Rc complex binding sites, within 200-300bp of the cap site, is of functional importance to steroid-inducibility. This multiplicity may prove to be the mechanism whereby S-Rc complexes identify specific genes otherwise lost in the enormous DNA excess presented by the eucaryotic genome. Several adjacent binding sites may act co-operatively, as is found for T-antigen binding to SV40 DNA in the control (repression) of early gene expression (reviewed Tjian, 1981).

In order to test these ideas it is desirable to have as many steroid/gene models as possible. Other groups have characterized excellent systems to study glucocorticoids, progestins and oestrogens as discussed above. However, to date there has not been a good model of androgenic steroid action, hence the attempt in this thesis to

characterize an androgen-responsive gene in the rat ventral prostate.

4. The Rat Ventral Prostate as a Model for the Study of Androgen Action.

Most of the glands within the male rat urogenital tract contribute secretory products to the seminal fluid. The seminal vesicle, ventral prostate, coagulating gland and epididymis all respond to androgenic steroids, resulting in tissue growth, cellular differentiation and increases in specific gene expression. Studies on the metabolism of androgens and their interaction with receptors have largely concentrated on the ventral prostate (Fang and Liao, 1971 and references therein) making it a well-characterized tissue, biochemically and physiologically. In normal animals, the ventral prostate contains many epithelial cells arranged in alveolar structures involved in the secretion of proteins for the ejaculate. Castration, resulting in testosterone withdrawal, leads to tissue regression and cessation of secretory activity. Secretion can be restored rapidly after an injection of testosterone (Price and Williams-Ashman, 1961) and cell numbers will also begin to increase after 24-36 hours with testosterone (Lesser and Bruchovsky, 1973).

The predominant constituent of the prostatic secretion is prostatic steroid binding protein (Heyns and De Moor, 1977) which has been independently identified by several groups and extensively characterized. The complete protein has an apparent Mr of 46,000 to 51,000, but readily dissociates into two subunits, S and F, of Mr 20,000 and 18,000 respectively. Upon reduction of disulphide bonds, S and F dissociate further to give a total of three distinct polypeptides: C3 (Mr 12,000 to 14,000) present in both S and F; C2 (Mr 11,000) from subunit S; C1 (Mr 8,000) from subunit F (Heyns *et al.*,1978a; Forsgren *et al.*, 1979). The ability of the protein to bind a wide variety of steroids with a high capacity and a low affinity, which is inversely proportional to the steroid polarity, has also been examined (Heyns and De Moor, 1977; Lea *et al.*, 1979). However, the actual function of the protein remains obscure. Nevertheless, prostatic steroid binding protein is an ideal marker of androgen action, as reviewed below.

In normal rats some 40% of the protein synthetic capacity of the ventral prostate is devoted to this protein. This level drops to 20%

24 hours after castration and declines to 2% after seven days. Injection of testosterone alone will reverse this decline in synthesis. The anti androgen, cyproterone acetate, will inhibit the response to testosterone and administration of oestrogens or glucocorticoid is ineffective (Parker et al., 1978).

To see if this effect on protein synthesis reflected an influence by androgens at the level of transcription, the base sequence complexity of $p(A^+)$ RNA from normal and castrated rats was analysed (Parker and Mainwaring, 1977; 1978). In normal rats an 'abundant class' of transcripts was observed containing 2-3 distinct sequences. However, in castrated animals an 'abundant class' was not found, showing that androgens could regulate the abundance of specific RNA transcripts within the ventral prostate. The cDNA to the 'abundant class' from normal prostate was isolated (Parker and Scrace, 1978) and shown to encode prostatic steroid binding protein by hybrid arrest of translation (Parker and Scrace, 1979). These observations were extended by analysing levels of individual mRNA's, achieved by cloning cDNA corresponding to mRNA to each of C1, C2 and C3, for use as specific probes (Parker et al., 1980). All three mRNA's declined from $\sim 10^5$ copies/cell to 10^2-10^3 copies/cell on castration. Administration of testosterone caused a fairly rapid increase in mRNA levels. After a short lag of ~1 hour, a 3-fold stimulation was observed within a further hour, independently of any cell proliferation (Parker et al., 1980). Therefore, androgens are able to increase prostatic steroid binding protein steady-state mRNA levels by 100-1000 fold. Subsequently it was shown that this increase is attributable in part to direct effects of testosterone on transcription of the genes. Studies on prostate nuclei isolated from normal and hormonally manipulated rats demonstrated a 2-3 fold effect of androgens on transcription of C1, C2 and C3. This means that androgens must also stabilize these mRNA's some 30-fold, probably at the level of a nuclear event (Page and Parker, 1982).

In order to provide some insight into how this androgenic control is achieved this thesis concentrates on the structure and expression of the genes to the C3 polypeptide. Being the most abundant of the three, this polypeptide is the best characterized, its amino acid sequence having been determined first (Peeters et al., 1981).

5. Interspersed Repetitive DNA.

During the course of this thesis, an Alu-like repetitive element was found very close to the 5' end of the C3 genes. A possible functional role for this element in the control of C3 expression was investigated. Therefore, for completeness, a review of the present knowledge on interspersed repetitive DNA is presented here.

Based on reassociation kinetics, repetitive DNA has been broadly classified into i) middle repetitive, present at 10^3-10^5 copies per haploid genome, and ii) highly repetitive, occurring at $\sim 10^6$ copies per haploid genome. Highly repetitive DNA (e.g. satellite DNA) is usually organized in clustered repeats and may play a role in chromosome structure and function, but is unlikely to influence protein coding genes (reviewed Brutlag, 1980). In contrast, middle repetitive elements are interspersed amongst single copy sequences, hence are in a position to influence gene expression, as proposed by Davidson and Britten (1979). This distribution of middle repetitive DNA is termed a short period dispersion pattern. In the rat this consists of repeated sequences of ~200-400bp interspersed among single copy DNA of 1.8 to 2.5kbp throughout 70 to 90% of the genome (Pearson et al., 1978; Wilkes et al., 1978).

The middle repetitive sequences of each species can be divided into a finite number of families based on sequence homology, usually assessed by cross-hybridization. For the rat, it has recently been estimated that there are ~60 such families (Witney et al., 1983). In general, members of one family differ from each other in sequence by 10-20%, but maintain a common structure or consensus sequence. The best characterized middle repeat family is that of the Alu elements of primates and rodents, so-called owing to the presence of a restriction site for Alu I within each repeat (Houck et al., 1979). The remarkable ubiquity of these sequences means that cloned isolates of most genes from these species contain Alu elements in their flanking DNA or introns (reviewed Jelinek and Schmid, 1982). This makes the study of Alu repeats particularly intriguing.

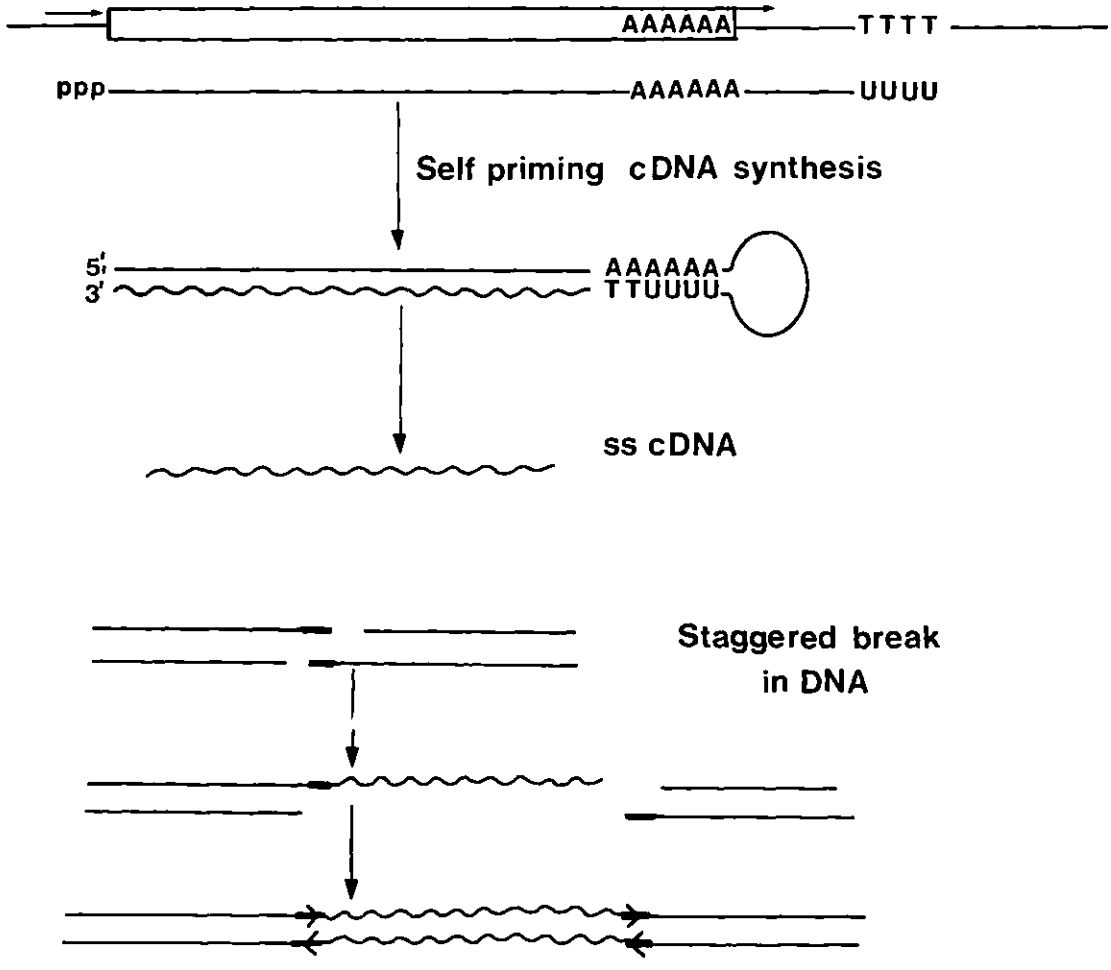
Several examples of human and rodent Alu elements have been cloned and sequenced and a picture of their structure and relatedness compiled (Jelinek and Schmid, 1982). Human Alu sequences are ~300bp long comprising an imperfect dimer of 130bp units with a 31bp insertion in the second monomer. Repeats found in rodent DNA are only 130bp long, but are homologous to the second monomer of the human Alu repeat,

except that they carry a rodent specific 32bp insertion instead of the 31bp sequence of humans. Collectively, these repeats are now called type 1 Alu sequences to distinguish them from the rodent type 2 Alu equivalent family characterized more recently (Haynes and Jelinek, 1981). Type 2 sequences are ~190bp long. The first 61bp show strong homology to residues 47 to 107 of type 1 sequences, but the remainder is type 2-specific (Haynes and Jelinek, 1981). Despite these differences, the two Alu families show the same type of structural organization: i) a precisely defined 5' end characterized by a consensus sequence for each family; ii) an imprecisely defined 3' end consisting of an A-rich sequence which, while being homologous among all Alu elements, is not identical even between members of the same family; iii) Alu sequences of all origins are flanked by direct repeats of 6-21bp within the cell DNA. These are unique in sequence for each Alu element. For this reason Alu repeats are believed to be mobile in the genome (but not transposable elements), the direct repeats being the product of each individual insertion event.

Alu elements can be transcribed into RNA by either RNA polymerase II or III. As Alu elements can occur in gene introns and 3' flanking DNA they may appear in mRNA transcribed by RNA polymerase II (Jelinek et al., 1980). Alternatively, human type 1 (first monomer only) and rodent type 2 sequences can be transcribed in vitro (Duncan et al., 1981) and in vivo (Haynes and Jelinek, 1981) by RNA polymerase III. These repetitive elements contain sequences bearing excellent homology with the consensus sequence for the bipartite polymerase III 'promoter' (Galli et al., 1981) and have been shown to be functionally equivalent (Paolella et al., 1983). The character of these Alu RNA's is also reminiscent of pol III transcripts with the 5' end of the RNA mapping to the start of the repeat and the 3' end mapping to a cluster of T residues (usually in single copy DNA beyond the repeat) resulting in a 3' oligoU sequence (Elder et al., 1981; see Figure 1). In many cases several termination regions are possible, producing families of transcripts with a common 5' end, but differing 3' sequences (Shen and Maniatis, 1982). So far, none of the proposed transcripts seems capable of being translated. The character of these Alu transcripts has lead to a proposal to account for the apparent mobility in the genome of Alu elements illustrated in Fig. 1. The mechanism suggested requires the return of RNA to the genome and has also been used to account for the structure of pseudogenes for

Figure 1. Proposed scheme to explain the apparent mobility of Alu elements within the genome.

The top line shows an Alu element, flanked by direct repeats (⟶) within single copy DNA, together with a possible RNA polymerase III transcript. Subsequent lines illustrate how the transcript may return to the genome via a self-primed DNA copy using reverse transcriptase, possibly provided by a retrovirus. This cDNA molecule may then integrate within genomic DNA at a staggered break, generated perhaps by two nearby topoisomerase I enzymes that have been made single-stranded breaks on opposite DNA strands. These enzymes can act at variable distances apart, giving different lengths of stagger which, when repaired, could result in the observed variable direct repeat lengths (6-21 bp) flanking inserted Alu sequences. These ideas and diagrams were combined from Jagadeeswaren et al. (1981), van Arsdell et al. (1981) and Bernstein et al. (1983).



Insertion of cDNA into the break and repair synthesis causes direct duplication of the DNA sequences at the break.

small nuclear RNA's (snRNA) (Bernstein et al., 1983) and for some protein coding genes (reviewed Sharp, 1983). Obviously the sequence of events proposed must occur in the germ line for the new insert to persist in the population.

Theories as to the role played by Alu elements and their transcripts (and middle repetitive RNA in general) are even more speculative. In some species, the pattern of repetitive DNA expression has been observed to change during development (reviewed Davidson and Posakony, 1982). Also families of small RNA's (4.5S, 4.5_{I}S, 7S and snRNA's; reviewed Zieve, 1981), believed to have functional roles in metabolism, have been found to show sequence homology to Alu elements (reviewed Jelinek and Haynes, 1982). Thus, type 2 transcripts may be involved in splicing (Georgiev et al., 1982) as proposed for U1 snRNA (Mount et al., 1983) or have a role in transport of $p(A^{+})$ RNA as has been suggested for 4.5_{I}S RNA (Leinwand et al., 1982). The 7SL RNA that forms an integral part of the signal recognition particle essential for protein secretion, also contains Alu-like sequences (Walter and Blobel, 1982). Other functions suggested for Alu elements include provision of gene transcription termination signals (Georgiev et al., 1982) and DNA replication origins (Jelinek et al., 1980).

In the absence of hard evidence as to the role played by middle repetitive DNA, it has been proposed that these elements are selfish. Their structure and transcription mechanisms may only serve to maintain these sequences within the genome, as an ultimate parasite (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). However, the mobility of these sequences within the genome can inadvertently lead to local DNA rearrangements which may play an accidental role in evolution (Calabretta et al., 1982).

The only way to determine whether middle repetitive DNA has a metabolic function though is to assess the importance of individual repeats adjacent to particular genes. Initially this must be done by correlating expression with a defined function as was attempted in the particular case of the C3 genes.

SECTION R: RESULTS

Objectives:

The overall aim of the work reported in this thesis was to characterise the gene(s) to the C3 polypeptide of the rat prostatic steroid binding protein with a view to establishing a model to study the mechanism of androgen-controlled gene expression.

Work on other steroid hormone inducible genes has shown the benefit of adopting a molecular approach. Therefore, it was necessary to isolate an active C3 gene from a rat DNA λ library which could be characterized in terms of its structure and in vivo pattern of expression via DNA sequencing and mapping of C3 mRNA to the gene.

When this project was commenced, incomplete cDNA clones for C1, C2 and C3 had been isolated in pAT 153 (Parker et al., 1980). These were used to screen a partial Eco RI λ library of rat DNA and several genomic clones had been isolated. Later the work was aided by the isolation and sequencing of a virtually full-length C3 cDNA clone, pA34, by other members of the laboratory (Parker et al., 1983).

In parallel with this project, the C3 genes were introduced into the mouse mammary carcinoma (S115) cell line, which contains androgen receptors. In many of the stably transformed cell lines isolated, androgen inducibility of C3 expression was retained (Page and Parker, 1983). This therefore illustrates one approach to the study of androgen-controlled gene expression made possible by the characterization of the C3 'model system'.

R1: ISOLATION AND CHARACTERIZATION OF C3 GENOMIC CLONES

1.1 Introduction

Preliminary screening of a rat DNA λ-library allowed the isolation of three identical C3 genomic clones in which the coding sequences were contained within a single Eco RI restriction fragment of 13.5 kbp. Unfortunately though these cannot represent all the C3 sequences within the rat genome. Eco RI digests of Sprague-Dawley rat DNA reveal, on probing for C3 sequences, up to four hybridizing bands which run as two doublets of 10.5/11.5 kbp and 12.5/13.5 kbp (see Fig. 4, track 1). Furthermore, not all individuals show all four hybridizing bands, as some lack the lower doublet while others lack one band from one doublet (Parker et al., 1983). These different hybridizing bands may therefore comprise either i) parts of a single gene, ii) allelic forms of one gene, with heterozygosity in the flanking sequences producing different sized restriction fragments, or iii) several copies of the C3 gene per haploid genome such that C3 is encoded by a multigene family.

Consequently, the initial aim was to isolate the remaining C3 coding sequences present in rat DNA. To this end the partial Eco RI library was rescreened together with a partial Hae III library, both constructed and amplified in λ Charon 4A from liver DNA isolated from a single male Sprague-Dawley rat (Sargent et al., 1979). The Hae III digest was cloned by the use of Eco RI linkers, thus the inserts could be separated from vector DNA by Eco RI digestion. As Hae III cuts more frequently in DNA than Eco RI, it was hoped that this second library would be more representative of the rat genome.

An additional library was also screened. This had been constructed in λgtWES.λB using a complete Eco RI digest of ventral prostate DNA from an individual of our Sprague-Dawley rat colony (Parker et al., 1983). The individual used had previously been shown to possess four distinct Eco RI fragments containing C3 sequences (Fig. 4, track 1).

1.2 Screening Methods

All three libraries were screened using published protocols with a few modifications, particularly the use of top agarose rather than top agar on the bacterial plates, which enabled nitrocellulose lifts to be taken cleanly. To ensure that the hybridization conditions favoured detection of C3 genomic clones, positive control filters carrying a previously isolated C3 clone, λ11B were included. The

brightest hybridization signals were achieved at temperatures of 62-65°C o/n using Schleicher and Schüll nitrocellulose.

The Charon 4A λ-libraries were initially probed with $[^{32}P]$-cDNA to total ventral prostate p(A^+) RNA so that genomic clones for C1, C2 and C3 could be isolated. This also circumvented the problem of the specific cDNA clones then available not being full-length and also that the partial Hae III library was reportedly contaminated with plasmid sequences. Once the λ-clones were 100% pure they were probed with specific $[^{32}P]$-nick translated cDNA clones to identify them. The λgtWES.λB library, however, was screened throughout using $[^{32}P]$-nick translated pA34, the full-length C3 cDNA clone.

1.3 Characterization of Recombinant Clones in λ Charon 4A

Primary screens of the partial Hae III and partial Eco RI libraries covered 6×10^5 and 3×10^5 plaques from each library, respectively. Twelve pure genomic clones were eventually isolated, each of which hybridized with one specific cDNA probe, as follows:

1 contained C1 genomic sequences (Hae III)
1 contained C2 genomic sequences (Eco RI)
10 contained C3 genomic sequences (3 Hae III/7 Eco RI)

The C1 and C2 clones resembled those already isolated and will not be discussed further. Nine of the ten C3 clones also contained essentially the same sequences isolated previously, giving the same pattern of hybridizing restriction fragments using a battery of different restriction enzymes. The most extensively studied of these new clones was the Hae III isolate, λ5A.

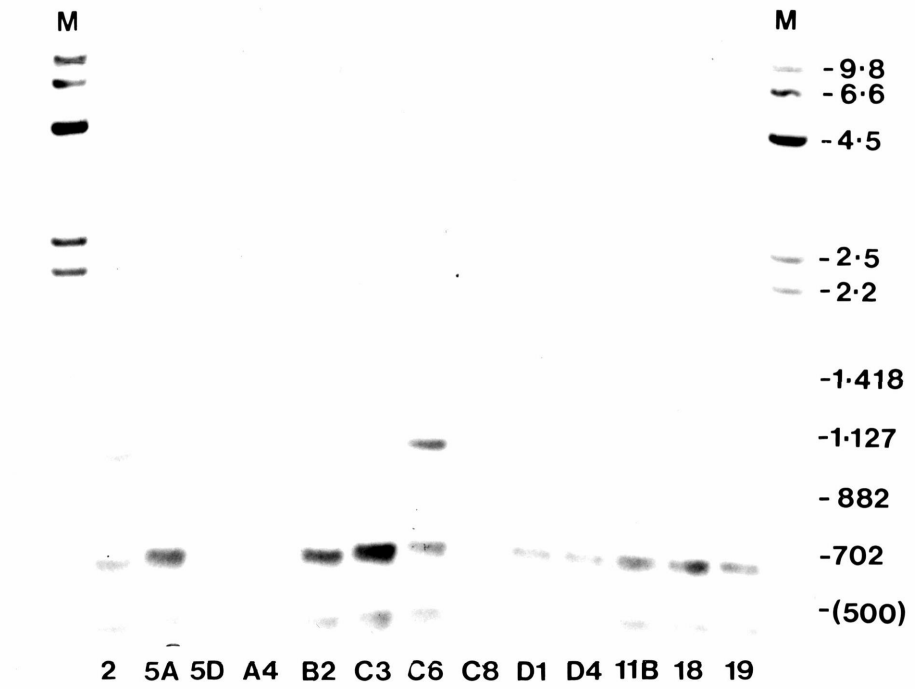
The tenth new clone, λ5D a Hae III isolate, appeared to be quite different from the others, giving multiple small bands on digestion with Eco RI. To check the relationship between λ5D and the other C3 clones, all ten new clones, plus three already isolated, were digested with Hae III and Eco RI, the products run out on a gel, blotted and probed for C3 sequences. As shown in Fig. 2, all the clones possessed four hybridizing bands except λ5D with only one. This implied λ5D was a partial C3 clone generated by digestion at an internal Hae III site. Further restriction mapping and blotting of the clone, shown in Fig. 3, lead to the conclusion that λ5D overlaps the 3' end of the inserts within the other C3 isolates and extends some 5kb beyond the

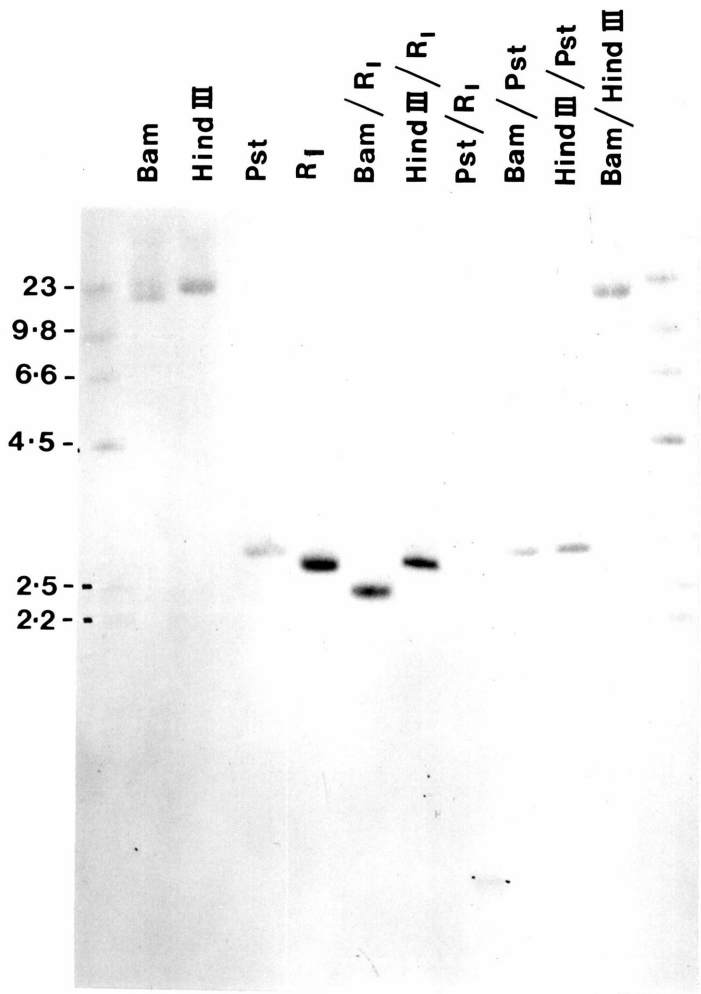
Figure 2. Analysis of 13 isolates from λ Charon 4A rat DNA libraries.

DNA (1 μg) from each clone was digested with Hae III and Eco RI and run o/n on a neutral 1% agarose gel. The gel was blotted and the nitrocellulose filter probed with $[^{32}P]$-cDNA to specific C3 mRNA which had been selected from prostate $p(A^+)$ RNA by hybridization to immobilized cloned C3 cDNA (Parker et al., 1980). The figure shows an o/n exposure of the washed filter. Clones 2, 5A and 5D were isolated from the partial Hae III library while the others were derived from the partial Eco RI library. The markers were 5' end-labelled Hind III digested λ and Hpa II digested polyoma DNA.

Figure 3. Restriction enzyme mapping of λ5D.

Each track shows 1 μg of cloned DNA digested with the indicated enzyme(s) and run o/n on a 0.8% neutral agarose gel. The gel was blotted and the filter probed as in Fig. 2. An o/n exposure of the washed filter is shown. The markers were 5' end-labelled Hind III digested λDNA.


M
M
-9·8
-6·6
-4·5
-2·5
-2·2
-1·418
-1·127
-882
-702
-(500)
2 5A 5D A4 B2 C3 C6 C8 D1 D4 11B 18 19


Bam
Hind III
Pst
R I
Bam / R I
Hind III / R I
Pst / R I
Bam / Pst
Hind III / Pst
Bam / Hind III
23 -
9·8 -
6·6 -
4·5 -
2·5 -
2·2 -

DNA isolated within these clones, shown diagrammatically in Fig. 6.

1.4 Characterization of C3 Clones Isolated in λgtWES.λB

The inability to isolate further C3 sequences from the λ Charon 4A libraries may have been due to genetic differences between the individual used to generate these libraries and our rat colony. Therefore, the λgtWES.λB complete Eco RI library was examined. This library (~3.5 x 10^5 plaques) was screened directly after packaging, without prior amplification, to ensure clones containing C3 sequences were not lost from the library. Two recombinant C3 clones, λ6 and λ8, were purified from the complete Eco RI library and restriction digests of their DNA were compared to similar digests of rat ventral prostate DNA (Fig. 4). Probing for C3 sequences showed that, while these two clones do not represent all the C3 coding regions present in cell DNA, they are distinct from one another. The complete Eco RI inserts of these two λ-clones were subcloned into pAT 153, as p61 and p81 respectively (see Appendix). Plasmid DNA was then used in a series of restriction digest and Southern blotting experiments, a selection of which are shown in Fig. 5.

λ8 proved to be identical to other C3 isolates as exemplified by λ11B and λ5A. However, λ6, while showing many common restriction sites, is clearly distinct, note for example the Pst I and Hind III digests. The complete λ6 Eco RI insert is also slightly smaller, being 12.5 kbp as against 13.5 kbp for the other clones. Restriction enzyme maps derived for all the C3 λ-clones are shown in Fig. 6, together with a map for the C3 cDNA clone, pA34. The orientation of the λ-clones was determined using specific 5' and 3' probes prepared from pA34. The clone was digested with Pst I, the 5' (160 bp) and 3' (450 bp) fragments isolated, radiolabelled by nick-translation and used separately to probe Southern blots of genomic clone digests. Such blots also helped to establish that the C3 coding potential of each genomic clone was split between three exons. Certain restriction enzymes used singly or in pairs allowed the coding regions to be divided into separate hybridizing bands, e.g. Msp I/Eco RI and Xba I/ Eco RI digests (Fig. 5, tracks 13 and 14; 15 and 16).

Figure 4. Comparison of C3 sequences isolated as λ6 and λ8 with those present in rat cell DNA.

The cell DNA tracks (C) contained rat ventral prostate DNA (20 μg) from one individual which was also used to generate the complete Eco RI library in λgtWES.λB (Parker et al., 1983). The other tracks contained 1 μg of cloned DNA from λ6(6) or λ8(8). All DNA samples were digested with restriction enzymes as indicated and run o/n on a 0.5% neutral agarose gel. Markers were 5' end-labelled Hind III digested λ DNA. The gel was blotted and the filter probed with [^{32}P]-nick-translated pA34 Pst I inserts using the dextran sulphate method (Page and Parker, 1982) in which 10% dextran sulphate was added to the hybridization buffer. A 3 day exposure of the filter is shown.

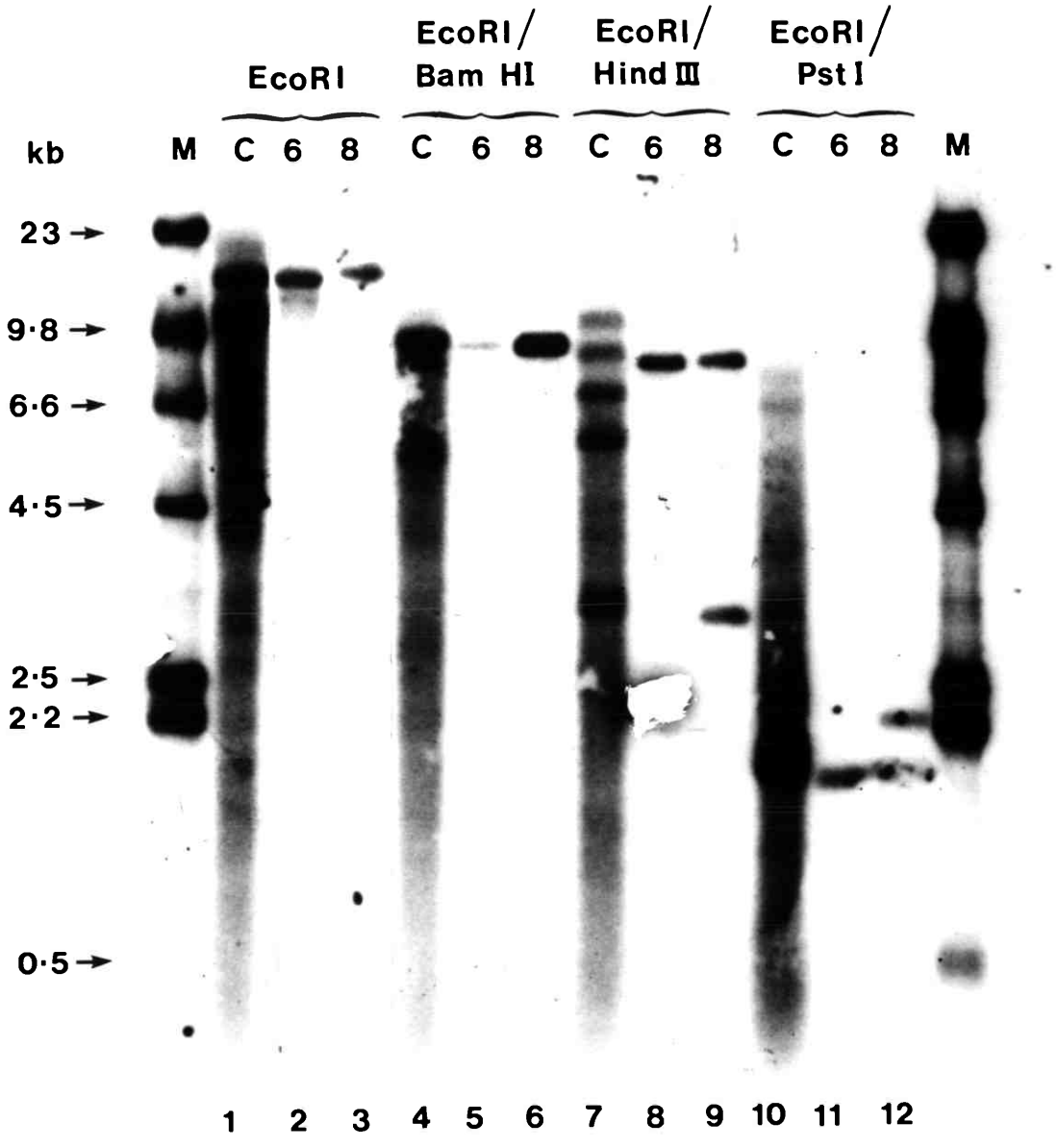

EcoRI
EcoRI/ Bam HI
EcoRI/ Hind III
EcoRI/ Pst I
kb
M C 6 8 C 6 8 C 6 8 C 6 8 M
23
9·8
6·6
4·5
2·5
2·2
0·5
1 2 3 4 5 6 7 8 9 10 11 12

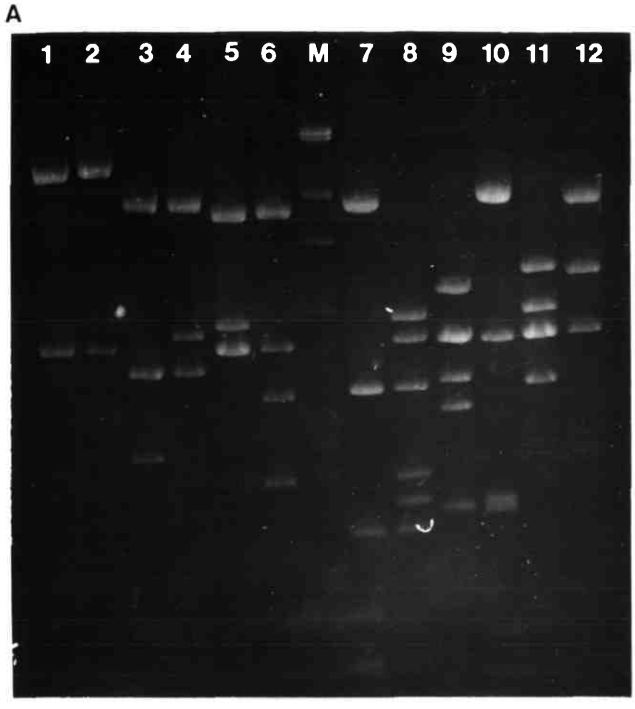
Figure 5. Comparative restriction enzyme digests of p61 and p81.

Odd numbered tracks represent p61 digests while even numbered ones show p81 digests. Each sample contained 1 µg DNA and all were digested with Eco RI plus other enzymes as listed below.

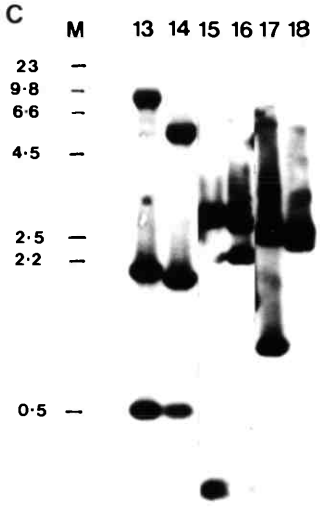
(A) A 0.7% neutral agarose gel of 12 comparative digests was blotted and probed with [^{32}P]-nick-translated pA34 Pst I inserts and the resulting autoradiograph is shown in (B). The digests were: tracks 1 and 2, Eco RI; tracks 3 and 4, Bam HI; tracks 5 and 6, Hind III; tracks 7 and 8, Pst I; tracks 9 and 10, Bgl II; tracks 11 and 12, Bst E_{II}.

(C) and (D) are exposures of filters probed as above showing further digests run on 0.8% neutral agarose gels as follows: tracks 13 and 14, Msp 1; tracks 15 and 16, Xba I; tracks 17 and 18, Hind II; tracks 19 and 20, Xba I with Bst E_{II}; tracks 21 and 22, Xba I with Pst I; tracks 23 and 24, Xba I with Hind II.

The markers used throughout were 5' end-labelled Hind III digested λ with 5' end-labelled Hinf I digested pAT 153 in D only.


A
1 2 3 4 5 6 M 7 8 9 10 11 12


B
1 2 3 4 5 6 M 7 8 9 10 11 12


C
M 13 14 15 16 17 18
23
9·8
6·6
4·5
2·5
2·2
0·5

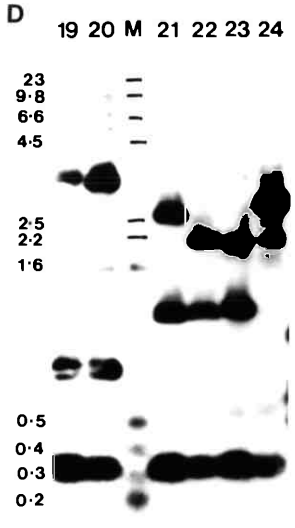
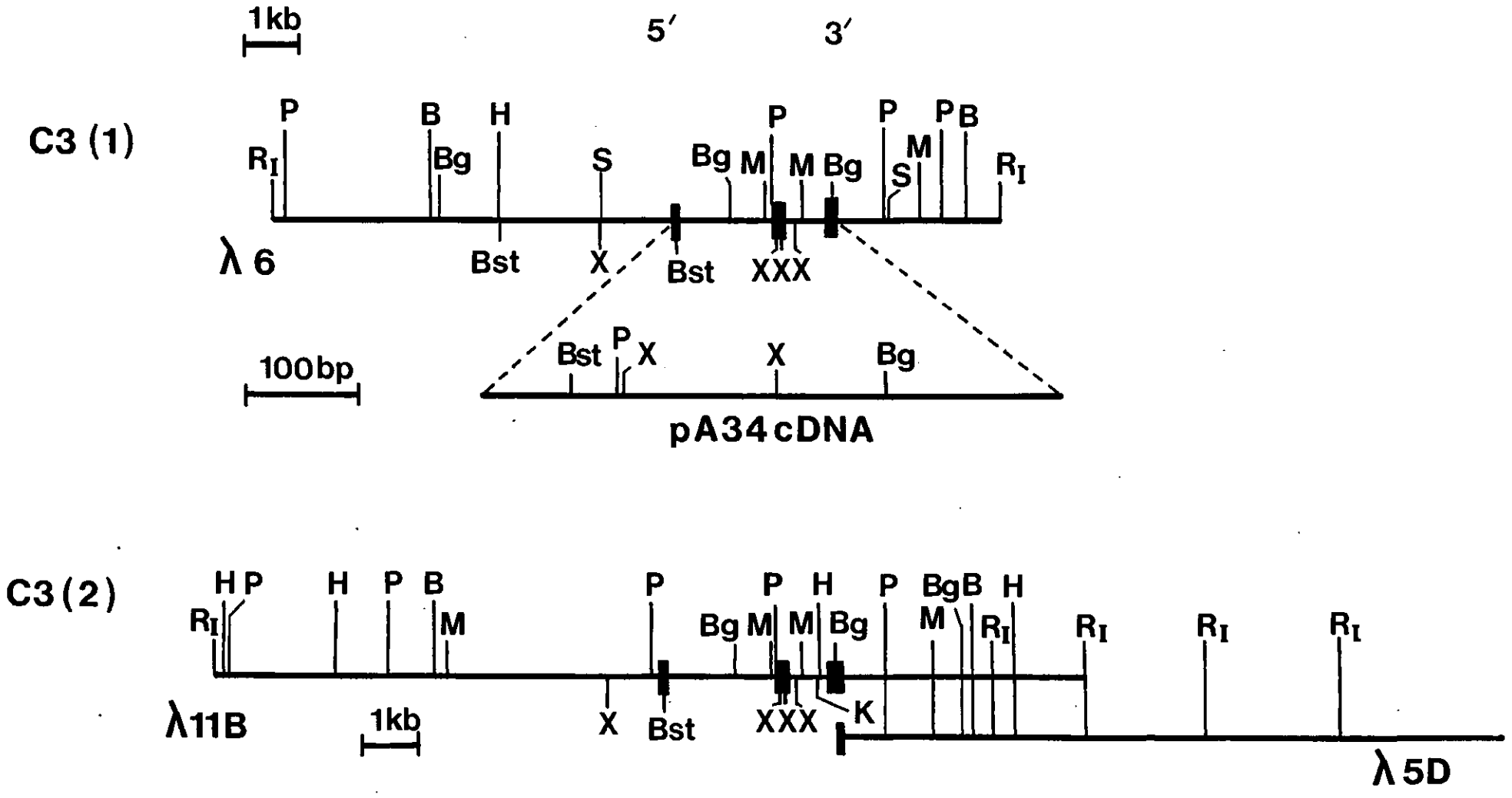

D
19 20 M 21 22 23 24
23
9·8
6·6
4·5
2·5
2·2
1·6
0·5
0·4
0·3
0·2

Figure 6. Restriction enzyme maps of the two C3 genes.

The map of C3(1) was deduced from restriction enzyme and R-loop analysis of λ6 and its pAT 153 subclones, while the C3(2) map was obtained from restriction enzyme digests of λ8, λ11B, λ5A and λ5D (and their subclones) together with R-loop analysis of λ5A. A restriction enzyme map of the C3 cDNA clone, pA34, is included for completeness. Solid blocks represent exons and the restriction enzyme sites indicated are as follows:-

B = Bam HI, Bg = Bgl II, Bst = Bst E_{II}, H = Hind III, K = Kpn I, M = Msp I, P = Pst I, RI = Eco RI, S = Sac I, X = Xba I.


C3 (1)
1kb
5′
3′
P
B
H
P
P
P
B
R_I
Bg
S
Bg
M
M
Bg
S
M
R_I
λ 6
Bst
X
Bst
XXX
100bp
Bst
P
X
X
Bg
pA34cDNA
C3(2)
H
P
H
P
B
P
P
H
P
Bg
B
H
R_I
M
Bg
M
M
Bg
M
R_I
R_I
R_I
R_I
λ11B
1kb
X
Bst
XXX
K
λ 5D

1.5 R-Loop Analysis of Genomic Clones

The three exon-two intron structure determined from Southern blotting for both types of genomic clone was examined further by R-loop analysis. Samples of λ6 and λ5A were hybridized with total $p(A^+)$ RNA and the RNA/DNA hybrids obtained examined by electron microscopy by Mrs. R. Tilley.

Examination of photographs of λ6 hybrids showed, as expected, three regions of RNA/DNA duplex shown in Fig. 7, A-D. The hybrid regions agreed well in length and separation with the exons mapped by Southern blotting, their relative orientation being deduced from the non-hybridizing 3' polyA tail of the mRNA. Averaged measurements from 28 R-loops gave exon sizes of 5' 0.13, 0.21 and 0.21 kbp 3' and intervening sequences as 1.72 ± 0.10 kbp and 0.77 ± 0.08 kbp. This implies C3 sequences of ~550 b total are encoded in the genome over 3.04 ± 0.18 kbp.

Analysis of R-loops formed with λ5A DNA did not give such a clear picture (Fig. 7, E-H). The first and third exons appeared to give RNA/DNA hybrids of 0.15 and 0.20 kbp respectively, but were separated by a region of duplex DNA of 2.9 kbp and a short 'bridge' of mRNA which did not hybridize together to delineate the middle exon, as observed for λ6. Apart from this, the corresponding exons are clearly of similar size from both genomic clones and the C3 coding sequences are spread over a similar length of DNA.

1.6 Discussion

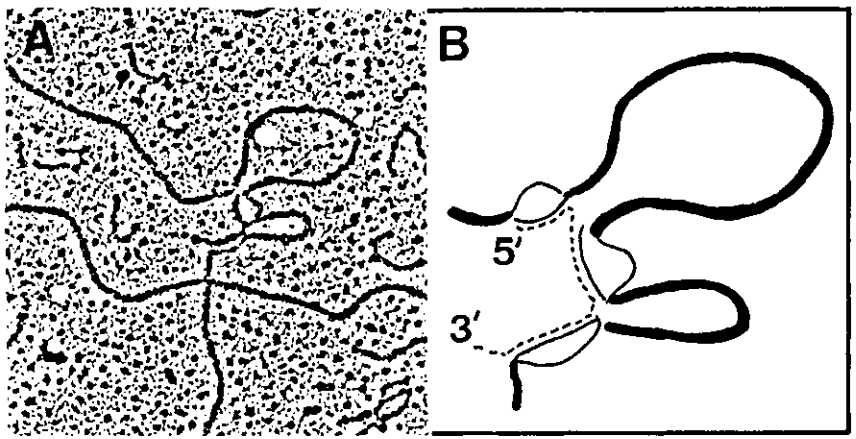
Repeated screening of three libraries has lead to the isolation of only one new type of C3 genomic clone as represented by λ6. The Eco RI insert of this clone is 12.5 kbp and is probably an isolate of one band in the upper doublet of a cell DNA Eco RI digest, as indicated in Fig. 4. The other band in this doublet has been isolated many times, λ5A, λ8, λ11B and (part of) λ5D being only a few examples.

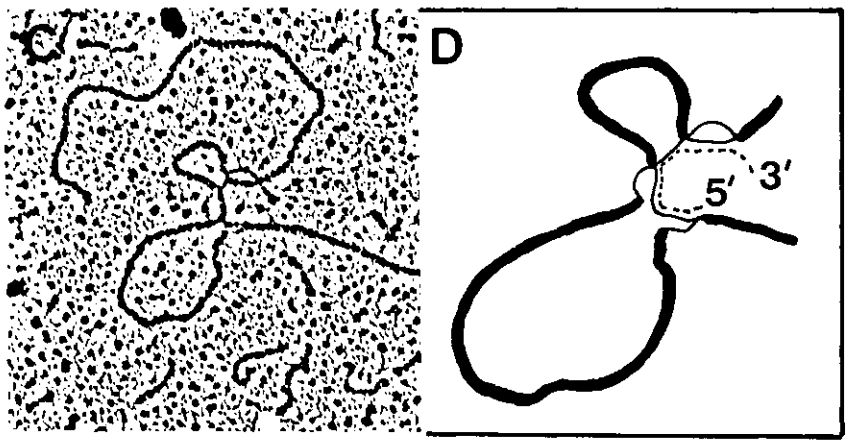
The reason for not isolating any C3 sequences represented in the 10.5/11.5 kbp doublet of Eco RI hybridizing bands remains obscure. A single copy restriction fragment of 10 kbp in a haploid genome of 3×10^9 bp would be represented at least at 1 copy in 300,000 within a λ-library, even without considering the selection imposed by the λ-vector itself. Thus, more than enough plaques were screened from all three libraries. Furthermore, λ Charon 4A and λgtWES.λB can be

Figure 7. Analysis of R-loops in the electron microscope.

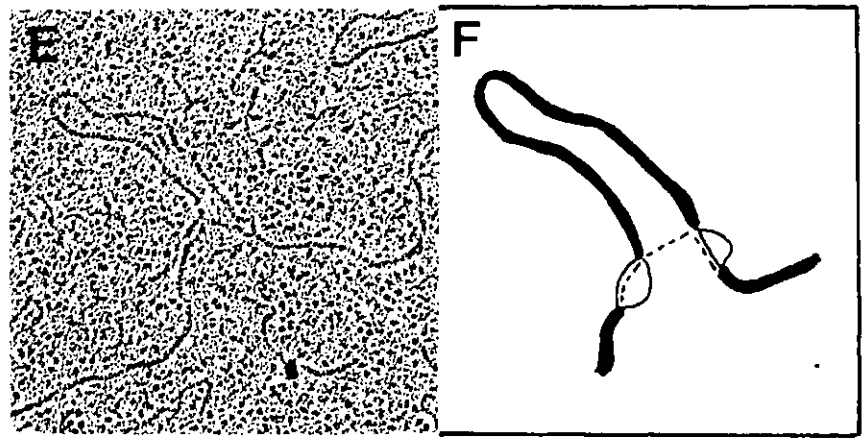
Electron micrographs (A, C, E and G) and diagrammatic representations (B, D, F and H) show R-loops formed between total ventral prostate p(A^+) RNA and λ6 (A-D) and λ5A (E-H). Thick, solid lines represent double stranded DNA, while thin lines represent single stranded DNA. RNA is shown by dashed lines.

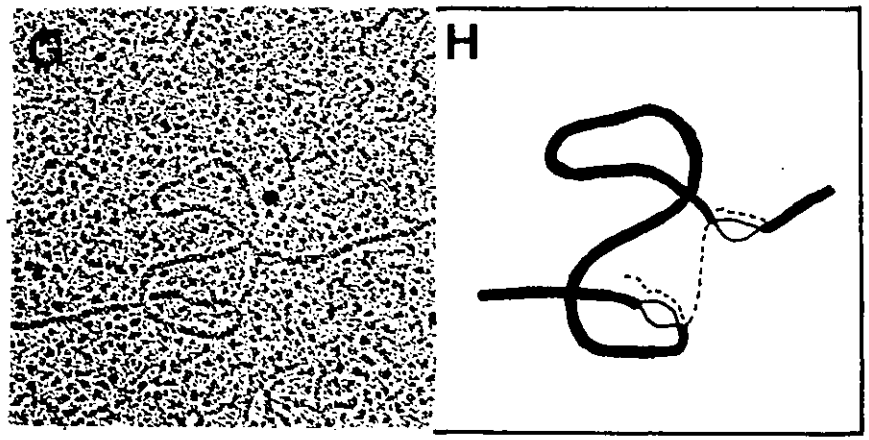
λ6





λ5A





efficiently packaged with inserts of 8.2-22.2 kbp and 4-14 kbp respectively, thus cloning of the 10.5/11.0 kbp doublet was well within the range of these vectors. Conceivably, the libraries made by Sargent's group used DNA from an individual not possessing an Eco RI 10.5/11.5 kbp doublet, or that certain C3 sequences were lost during amplification of these libraries. Such reasoning cannot be applied to the construction of our own, non-amplified library using DNA from a rat carrying all four Eco RI fragments. One possible explanation is that the DNA flanking the C3 coding sequences within these restriction fragments contains so-called poison sequences that inhibit cloning in λ. A similar observation has been made to account for the failure to isolate clones containing the 5' portion of mouse mammary tumour virus, using either exogenous viral or integrated proviral DNA (Groner et al., 1980; Donehower et al., 1981).

In view of the failure to isolate all the C3 sequences within cell DNA, it was essential to establish the relationship between the two types of C3 clone obtained. From the restriction mapping and R-looping data the two clones appear to represent two similar, but distinct complete C3 genes. Clearly both genes contain sufficient coding potential to produce the 509 b of C3 mRNA, isolated as the cDNA clone pA34 which has been shown, by sequencing (Parker et al., 1983), to be compatible with the C3 polypeptide sequence (Peeters et al., 1981). The idea that there are two C3 genes per haploid genome is also consistent with data from Eco RI digestion of cell DNA from individual rats (Parker et al., 1983). Those animals that lack the lower doublet are homozygous for the two C3 genes, while those possessing the 10.5/11.5 kbp doublet are heterozygous and show polymorphism at the Eco RI sites within the DNA flanking the genes. Consequently, the uncloned C3 sequences are probably alleles of the two genes isolated. However, the Sprague-Dawley rat strain is not highly in-bred and it is difficult to be wholly conclusive without conducting a series of breeding experiments. Therefore, it is conceivable that the four hybridizing Eco RI fragments represent four distinct C3 genes.

Given that two C3 genes have been isolated, it must be established that at least one is active in vivo. Although the two genes are apparently very similar, the middle exon of the gene cloned in λ11B, λ5A, λ8 etc. has difficulty hybridizing with C3 mRNA. This is shown in the R-looping study, where an RNA/DNA hybrid delineating the second exon was not observed for these clones and in some of the Southern

blotting experiments. Tracks 13-16 in Fig. 5, containing Msp I/ Eco RI and Xba I/Eco RI digests were mentioned earlier. In each case the smallest hybridizing fragment represents the middle exon and these bands always hybridize more strongly with nick-translated $[^{32}P]$-pA34 inserts in the p61 tracks than in the p81 tracks. Hence, these two very similar genes clearly show divergence, at least within the middle exon, with the λ6 gene being more likely to be expressed in vivo as it is more compatible with the C3 cDNA clone. For this reason the λ6 gene has been designated C3(1) while the gene isolated in λ11B etc. has been termed C3(2).

R2: DNA SEQUENCE ANALYSIS OF C3(1) AND C3(2)

2.1 Introduction

The DNA sequence of the two C3 genomic clones was determined fundamentally to establish that at least one of them could encode authentic C3 protein and therefore be likely to be active *in vivo*. Sequencing was also a prerequisite for full characterization of the C3 gene family in the hope of providing insights into the androgen control of gene expression.

As the C3 genes did not appear to possess sufficient, easily labelled restriction sites for successful Maxam and Gilbert chemical degradation sequencing, the sequencing methods pioneered by Sanger were preferred. These are based on the hybridization of a ss universal primer to cloned ss DNA and extension of the primer in four separate reactions, each containing the four dNTP's (one radiolabelled) and one ddNTP acting as a random chain terminator. The four reaction mixes are then run out on acrylamide gels to separate the primer extended fragments on a basis of length. They will have a common 5' end (the primer) and random 3' ends which, over the four reactions, will correspond to every nucleotide in the sequenced DNA (usually 250-300 b). The sequence may therefore be read directly from the gel autoradiograph.

Methods whereby suitable cloned ss DNA is obtained for this purpose rely heavily on the biology of the ss bacteriophage, M13. This is a filamentous phage with a closed circular genome of 6.5 kbp (Denhardt *et al.*, 1978). Since the phage infect *E.coli* through the sex pilus they can only replicate in male bacteria, carrying the F episome. During the replicative cycle the phage initially becomes double stranded and this replicative form (RF) accumulates to ~100 copies per cell. Thereafter DNA synthesis becomes asymmetric, producing large amounts of the infecting (+) strand, each of which acquires a protein coat. The resulting phage particles are continually secreted from infected cells without producing lysis, though the infection does cause the bacteria to show retarded growth. Consequently the virus produces turbid plaques of slow growing bacteria on a bacterial 'lawn'. The large amounts of ss DNA, equivalent to only one strand of the virus, that can be isolated from the culture medium of infected bacteria are ideal for 'dideoxysequencing' making M13 an excellent cloning vector for this purpose. The ds RF can be isolated from infected bacteria for

the cloning of ds genomic fragments which can then be recovered as ss DNA clones from recombinant phage particles. However, large inserts are rather unstable in M13, so fragments of 300-1000 bp alone may be cloned and sequenced and a series of overlapping clones is needed to determine a complete gene sequence.

The M13 cloning vectors and bacterial hosts have been specifically modified so that M13 recombinant clones may be readily identified. The host bacterium, E.coli JM101 (Messing, 1979), lacks a chromosomal gene for β-galactosidase. Instead, the F episome carries a defective lac gene which produces an enzyme lacking amino acid residues 11-41 of the α-peptide. The M13 cloning vectors mp7, 8 and 9 (Messing et al., 1981; Messing and Vieira, 1982) also carry part of the lac gene sufficient to produce a complete α-peptide. Therefore, infection of an M13 vector into JM101 in the presence of the inducer, I.P.T.G., will produce two defective proteins which can complement each other to give full β-galactosidase activity. This is detectable using the colourless galactose analogue, B.C.I.G., which is hydrolysed by the enzyme to give a blue dye that will colour 'plaques' of infected bacteria.

The M13 vectors also carry a polylinker (containing several unique restriction sites for cloning) within the α-peptide coding region. While this does not inhibit α-complementation in the parent vector, the introduction of foreign DNA at this point in recombinant clones causes disruption of the coding region. This means α-complementation is blocked and white plaques are produced in the presence of I.P.T.G. and B.C.I.G., denoting recombinant clones that may be grown up to obtain ss DNA for sequencing. The sequencing primer hybridizes to a constant region downstream of the polylinker, hence can be used to sequence all the clones.

The polylinker in M13 mp7 contains pairs of restriction enzyme sites so is only suitable for cloning DNA fragments produced by the action of one enzyme. Such fragments will be cloned in either of two orientations and theoretically 50% of the resulting clones will be in each orientation. Since in practice one DNA strand is often cloned more readily than the other, the vectors mp8 and 9 are useful. They contain half the polylinker present in mp7 in opposite orientations so that not only fragments produced by one restriction enzyme digest may be cloned, but also fragments from double digests. Such fragments, cloned into suitable double digests of mp8 and 9, will automatically

produce clones of each DNA strand for efficient cloning and sequencing.

In applying these methods to the sequencing of the two C3 genes several problems were encountered, particularly in obtaining sufficient clones to cover the entire coding region.

2.2 Conditions for 'Dideoxysequencing'

Detailed protocols used in the sequencing reactions are given in Section M2. Preliminary experiments were carried out to optimize these reaction conditions and thereby improve the resolution of the final gel autoradiographs, as discussed below:

i) Primer concentration. The Biolabs 15b universal primer was used, which hybridizes 18 or 21 b downstream of the polylinker, depending on the vector used. The concentration of primer in each sequencing reaction recommended by Biolabs (50 pM) was found to be too high, producing many artefact bands. A series of dilutions of the primer were used to empirically determine the optimum concentration and this was found to be 7.5 pM (a dilution of 6.5 fold), though concentrations involving 5 to 8 fold dilutions all gave readable sequences.

ii) ddNTP concentration. The ratio of concentrations of ddNTP to dNTP used initially was 40:1, as recommended by Dr. Sanger's laboratory. This proved to be excellent for the A, G and T reactions, but was too high for the C reaction, giving excessive termination at early stages. The ddC:dC ratio was finally altered to 18:1 to remove this problem.

iii) Prechase sequencing. Using standard sequencing conditions it was possible to routinely sequence ~300 b from each M13 subclone, the limiting factor being the concentration of $[\alpha^{32}P]$-dCTP. Therefore, to obtain sequence beyond 300 b a cold prechase was adopted so that labelling and chain termination started later in the sequence and the limiting factor became the degree of resolution of the gels. Using 3 or 4 different gel running times it was possible to obtain 400-450 b of sequence.

iv) Other primers. Another method of obtaining more sequence was to make C3 specific primers from the cDNA clone, pA34. Unfortunately RNA added to precipitate the primers during their preparation caused a dark background on the sequencing gels, but some success was

achieved using an Xba I - Alu I, 68 b primer hybridized to 5' clones of C3(2) (see Figs. 13A and 14A).

2.3 Cloning and Sequencing Strategy

Subcloning of restriction fragments from both C3 genes into M13 vectors was greatly facilitated by the previous construction of subclones in pAT 153, whose structures are detailed in the Appendix. The overall strategy used to sequence C3(1) and C3(2) is shown diagrammatically in Figure 8 while particular points and problems are discussed below.

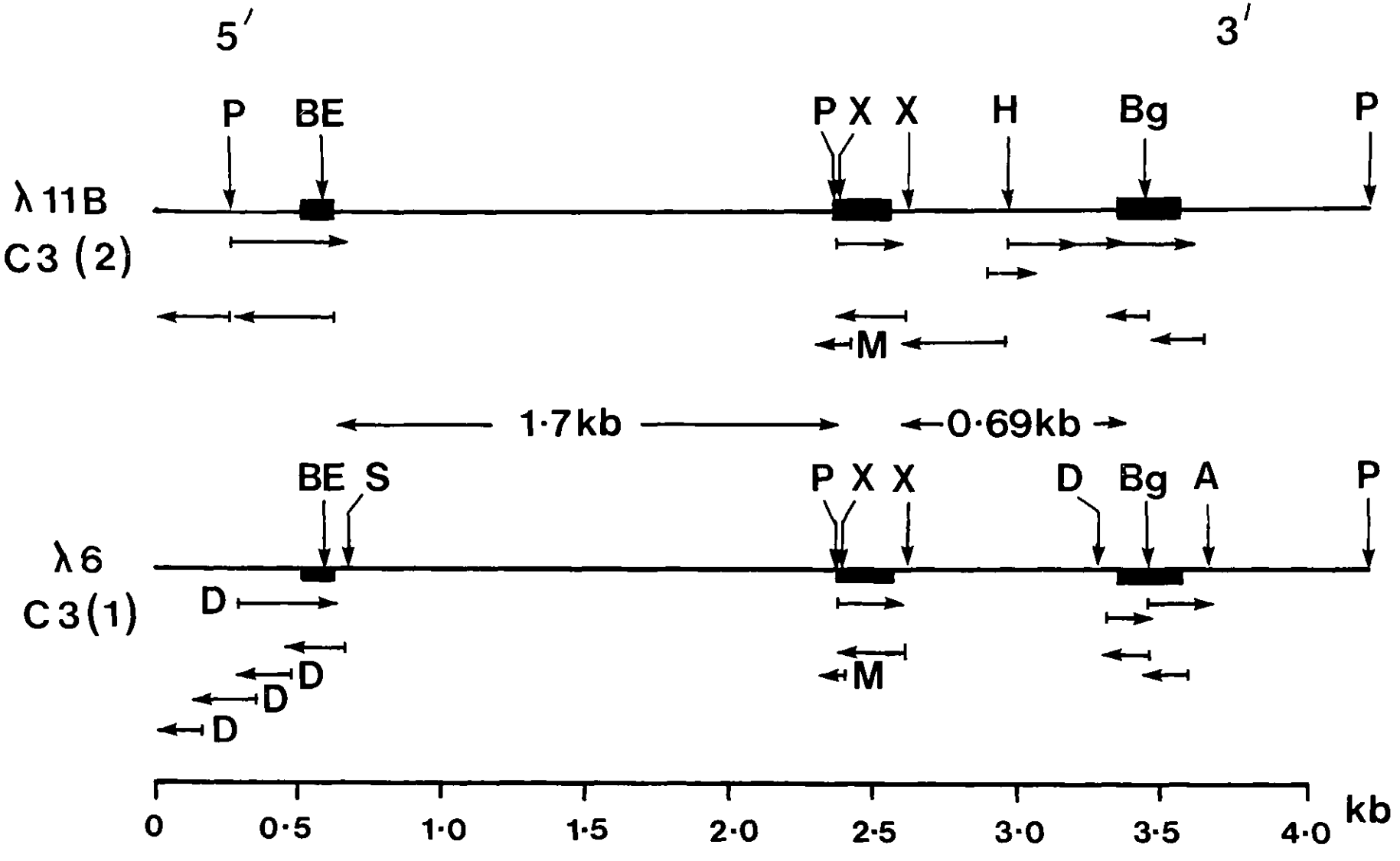
Initially, it was attempted to subclone C3(2) (the first genomic clone isolated) into M13 using a shotgun cloning procedure. A battery of restriction enzymes was tested to identify those that produced fragments of C3(2) of less than 1 kb. The enzymes Sau 3A, Hae III and Alu I were selected and individual digests of p11B1 (5' end) and p11B2 (3' end) were ligated into M13 mp7, digested with Bam HI or Hind II. It was hoped this would result in a library of short, overlapping, easily sequenced clones representing both DNA strands. Several clones containing coding regions were sequenced. These were identified by probing dot blots of the isolated M13 recombinants with $[^{32}P]$-nick-translated pA34 Pst I inserts. Unfortunately the library of clones was not very representative as some sequences did not clone at all while others were only obtained on one DNA strand. The 'blunt-end' ligations of Alu I and Hae III digests into the Hind II site of M13 were particularly difficult, producing only 10-20% of white plaques per plate as against 70-80% for 'sticky-end' ligations.

This approach was therefore abandoned in favour of a more direct one concentrating on obtaining clones (on both DNA strands) for each exon and the 5' and 3' flanking DNA. This was achieved by the isolation from pAT 153 subclones of specific DNA fragments that were recloned into mp8 and 9. These are detailed for C3(1) and C3(2) in the Appendix. This approach, too, was not without problems however.

Compared to C3(2), it was difficult to obtain clones of the first exon and 5' flanking DNA of C3(1) due to the absence of the 5' Pst I site within this gene. For this reason two series of C3(1) 5' deletion mutants were constructed as described in the Appendix and indicated by a 'D' in Fig. 8. While these clones were relatively easy to produce and propagate as Bam HI/Eco RI inserts in pAT 153,

Figure 8. DNA sequencing strategy.

The organization of the two C3 genes as represented by the clones λ11B and λ6 is presented, with exons marked as solid blocks. The restriction sites shown are: A = Alu I, BE = Bst E_{II}, Bg = Bgl II, D = Dde I, H = Hind III, P = Pst I, S = Sau 3A, X = Xba I. In each case only those sites actually used for cloning particular fragments (described in the Appendix) are given. Arrows indicate the positions of each DNA strand which has been sequenced, ⟶ are non-coding sequences; ⟵ are coding sequences. A 'D' shows that the sequence was derived from a deletion mutant, while an 'M' indicates those sequences determined by Maxam and Gilbert chemical degradation sequencing.


5′
3′
P
BE
P X X
H
Bg
P
λ11B
C3 (2)
M
1·7kb
0·69kb
BE S
P X X
D Bg A
P
λ6
C3(1)
D
D
D
D
M
0
0·5
1·0
1·5
2·0
2·5
3·0
3·5
4·0
kb

transferring the inserts to mp8 and 9 proved less successful. The A series of mutants (3' deletions) only cloned into mp8, giving the coding strand sequence and all attempts to clone into mp9 failed. The B series (5' deletions) were constructed to clone also into mp8, but to produce the non-coding strand sequence, however, M13 recombinants were only isolated using the shortest mutant, Δ3 (-195B). In view of these difficulties only the proximal ~200 bp of C3(1) 5' flanking sequence could be determined on both DNA strands. Interestingly, sequencing from both DNA strands of C3(2) upstream of the 5' Pst I site (at -255) also proved impossible, again as only the non-coding strand would clone. As all of these sequences proved to be AG-rich and very repetitious it was possible the inserts were unstable within M13 vectors.

Another problem observed for both C3(1) and C3(2) was the inability to obtain any M13 recombinants using DNA fragments that included the 3' end of the first intron, to the left of the central Pst I site. The host bacterium, JM101, is E.coli K12 restriction positive so sequences containing unmodified Eco K sites will be rarely cloned as the transfected recombinant DNA would be linearized and therefore not propagated in these bacteria. Several attempts were made to clone specific restriction fragments from this region using the host JM103 which is reportedly r_K^- (Messing et al., 1981), but these too failed and it now turns out that JM103 is r_K^+ (Baldwin et al., 1983).

Consequently, to obtain the sequence across the central Pst I site, including the first seven bases of the second exon, the methods of Maxam and Gilbert were used. A similar strategy was adopted for C3(1) and C3(2) using the plasmid subclones p62A and p5A2 respectively (see Appendix and Fig. 9). The sequence obtained included the suspected Eco K recognition site, which had caused all the M13 cloning difficulties, as shown in Fig. 9.

Eventually, using the cloning and sequencing strategies described, it was possible to sequence both strands of the coding regions and immediate 5' and 3' flanking DNA of C3(1) and C3(2). In addition, sequences upto -585 (for C3(2)) and -685 (for C3(1)) have been obtained for one strand (coding) only. A comparison of the two genomic sequences is shown in Fig. 10, while the available intervening sequence data is compared in Fig. 11. All of these sequences were determined at least twice.

Figure 9. Maxam and Gilbert DNA sequence analysis across the central Pst I site.

The C3(1) and C3(2) subclones p62A and p5A2 (see Appendix) were sequenced from the Hind III site within the plasmid DNA (hatched boxes) into the first intron of the gene as shown for p62A in the diagram (exons shown as solid blocks). DNA (5 µg) was 5' end-labelled at the Hind III site and redigested with Nae I. The purified 150 bp labelled fragment was then subjected to chemical degradation reactions (Maxam and Gilbert, 1980) as follows: A + G reaction 40 mins at 37°C; G reaction 7 mins, C reaction 12 mins and T + C reaction 13 mins, all at 22°C. The products of these reactions were run on a 6% acrylamide gel for 1½ hrs and the gel was autoradiographed for 3 days at -70°C using Kodak Noscreen film. The sequence reproduced at the side of the gel represents the C3(1) coding strand within which the Eco K recognition sequence is shown

5' G C A C$(N)_6$G T T 3'
3' C G T G$(N')_6$C A A 5'

DNA sequence immediately adjacent to the labelled nucleotide is usually difficult to ascertain, as the very short oligonucleotides run aberrantly on the gel. However, by labelling at a site 29 bp away, within the plasmid vector, it was possible here to obtain the full C3 sequence within this region of the subclone.

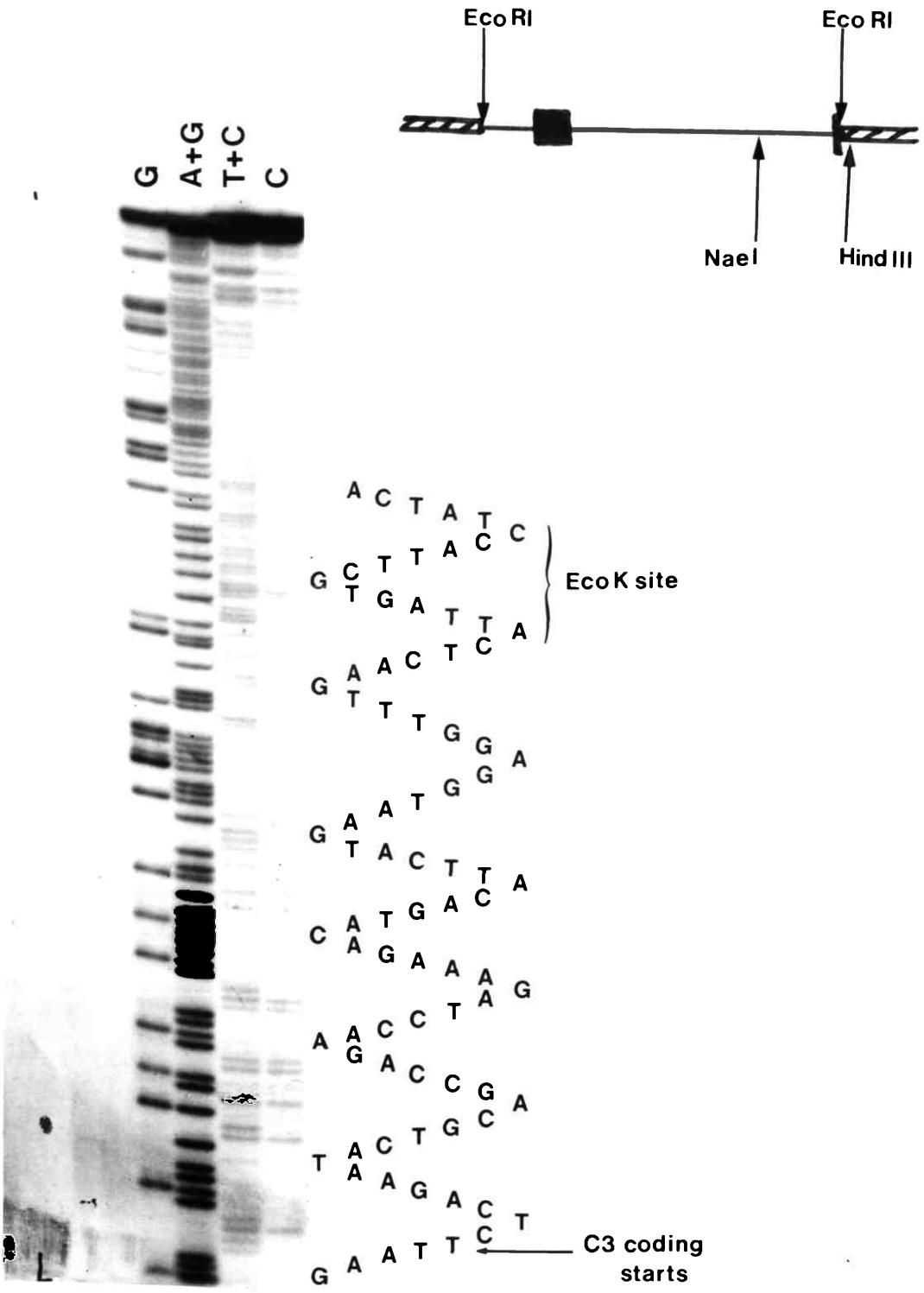

Eco RI
Eco RI
Nae I
Hind III
G
A+G
T+C
C
Eco K site
C3 coding starts

Figure 10. DNA sequence of two C3 genes.

The nucleotide sequence for the non-coding strand of the C3(1) gene, together with the corresponding amino acid sequence, is presented. Differences within the C3(2) gene are shown beneath the C3(1) sequence, including deletions (marked $\vee$) and insertions (marked $\wedge$). TATA boxes, splice acceptor and donor sites and the putative polyadenylation signal are boxed. Transcription regions T_1 and T_2 are labelled and the 3' transcription terminus marked with an asterisk (*). Vertical arrows ($\downarrow$) delineate the exact exon-intron junctions, though only sufficient intervening sequence to compare with the consensus splice donor $\left[\begin{matrix}C\\A\end{matrix}\ A\ G\downarrow G\ T\ \begin{matrix}G\\A\end{matrix}\ A\ G\ T\right]$ and acceptor $\left[\left(\begin{matrix}T\\C\end{matrix}\right)_n\ N\ \begin{matrix}C\\T\end{matrix}\ A\ G\downarrow G\right]$ sequences (Mount, 1982) is shown. The numbering corresponds with that given for pA34 (Parker et al., 1983) and does not include any intervening sequences.

The Alu-like repetitive element is underlined while possible examples of flanking direct repeats are overlined.

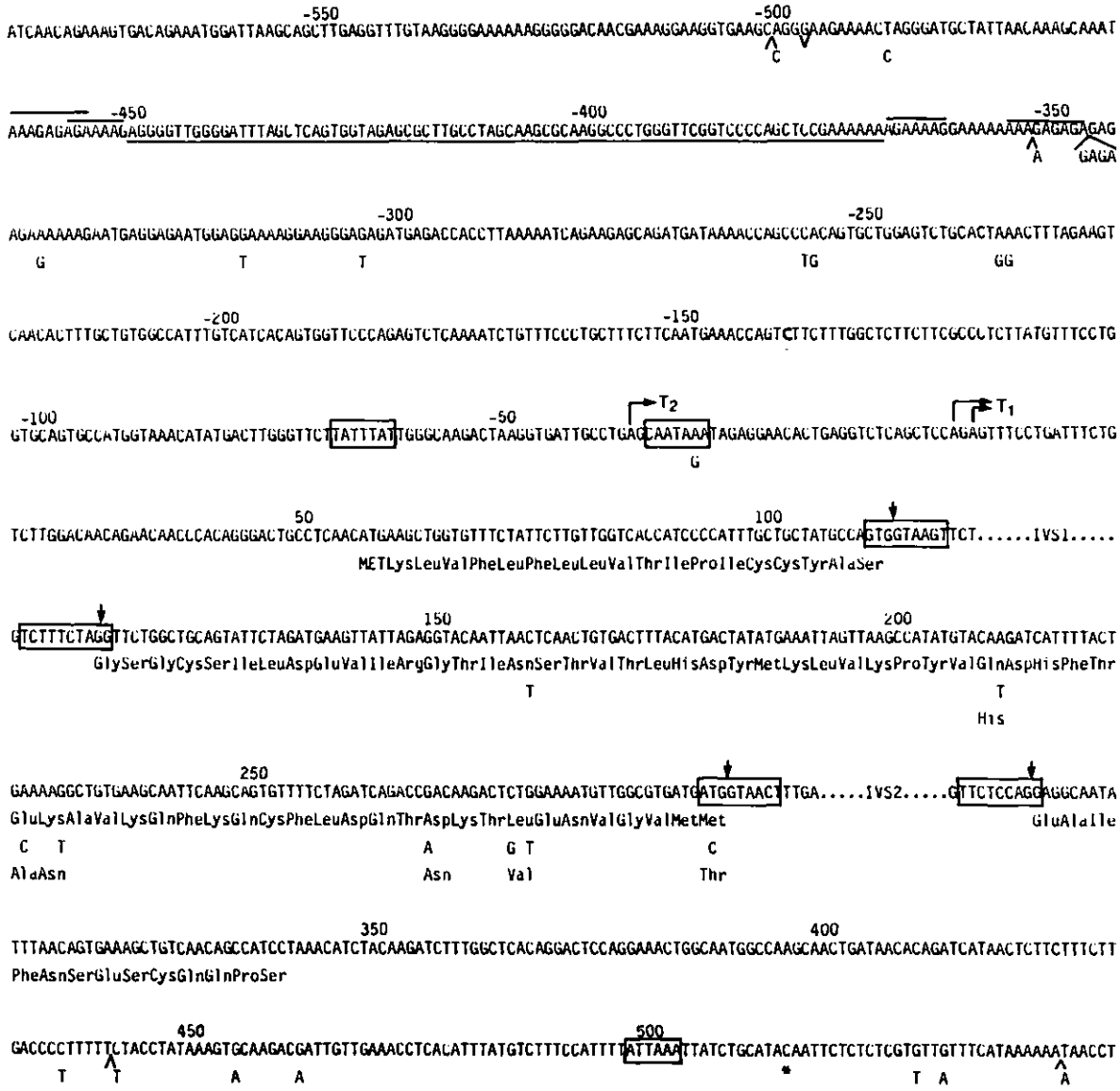

-550
-500
ATCAACAGAAAGTGACAGAAATGGATTAAGCAGCTTGAGGTTTGTAAGGGGAAAAAAGGGGGACAACGAAAGGAAGGTGAAGCAGGGAAGAAAACTAGGGATGCTATTAACAAAGCAAAT
C
C
-450
-400
-350
AAAGAGAGAAAAGAGGGGTTGGGGATTTAGCTCAGTGGTAGAGCGCTTGCCTAGCAAGCGCAAGGCCCTGGGTTCGGTCCCCAGCTCCGAAAAAAAGAAAAGGAAAAAAAAGAGAGAGAG
A
GAGA
-300
-250
AGAAAAAAGAATGAGGAGAATGGAGGAAAAGGAAGGGAGAGATGAGACCACCTTAAAAATCAGAAGAGCAGATGATAAAACCAGCCCACAGTGCTGGAGTCTGCACTAAACTTTAGAAGT
G
T
T
TG
GG
-200
-150
CAACACTTTGCTGTGGCCATTTGTCATCACAGTGGTTCCCAGAGTCTCAAAATCTGTTTCCCTGCTTTCTTCAATGAAACCAGTCTTCTTTGGCTCTTCTTCGCCCTCTTATGTTTCCTG
-100
-50
T2
T1
GTGCAGTGCCATGGTAAACATATGACTTGGGTTCTTATTTATTGGGCAAGACTAAGGTGATTGCCTGAGCAATAAATAGAGGAACACTGAGGTCTCAGCTCCAGAGTTTCCTGATTTCTG
G
50
100
TCTTGGACAACAGAACAACCCACAGGGACTGCCTCAACATGAAGCTGGTGTTTCTATTCTTGTTGGTCACCATCCCCATTTGCTGCTATGCCAGTGGTAAGTTCT.....IVS1.....
METLysLeuValPheLeuPheLeuLeuValThrIleProIleCysCysTyrAlaSer
150
200
GTCTTTCTAGGTTCTGGCTGCAGTATTCTAGATGAAGTTATTAGAGGTACAATTAACTCAACTGTGACTTTACATGACTATATGAAATTAGTTAAGCCATATGTACAAGATCATTTTACT
GlySerGlyCysSerIleLeuAspGluValIleArgGlyThrIleAsnSerThrValThrLeuHisAspTyrMetLysLeuValLysProTyrValGlnAspHisPheThr
T
T
His
250
GAAAAGGCTGTGAAGCAATTCAAGCAGTGTTTTCTAGATCAGACCGACAAGACTCTGGAAAATGTTGGCGTGATGATGGTAACTTTGA.....IVS2.....GTTCTCCAGGAGGCAATA
GluLysAlaValLysGlnPheLysGlnCysPheLeuAspGlnThrAspLysThrLeuGluAsnValGlyValMetMet
GluAlaIle
C T
A
G T
C
AlaAsn
Asn
Val
Thr
350
400
TTTAACAGTGAAAGCTGTCAACAGCCATCCTAAACATCTACAAGATCTTTGGCTCACAGGACTCCAGGAAACTGGCAATGGCCAAGCAACTGATAACACAGATCATAACTCTTCTTTCTT
PheAsnSerGluSerCysGlnGlnProSer
450
500
GACCCCTTTTTCTACCTATAAAGTGCAAGACGATTGTTGAAACCTCACATTTATGTCTTTCCATTTTATTAAATTATCTGCATACAATTCTCTCTCGTGTTGTTTCATAAAAAATAACCT
T
T
A
A
*
T A
A

Figure 11. Comparison of intervening sequences from the two C3 genes.

Only the second intron of C3(2) was completely sequenced, but all the intervening sequence data available from both genes are presented here comparing the non-coding strand sequences. The exon-intron boundaries are indicated using vertical arrows (↓) and the coding sequences show nucleotide numbers corresponding to those in Fig. 10. Differences between the two sequences are marked with asterisks (*).

1ST EXON/1ST INTRON BOUNDARY

```
                          100                                          *
C3(1)     TCACCATCCCCATTTGCTGCTATGCCAGTGGTAAGTTCTGTAGAAGGGAAGAGTG CTTT
C3(2)     TCACCATCCCCATTTGCTGCTATGCCAGTGGTAAGTTCTGTAGAAGGGAAGAGTGTCTTT

                         *          * **   *  *    **
C3(1)     TGGACCACACATTCTTACTGTGTCTTCCAGGAAGAAGTCATGATACCTGAGCACTAGTAT
C3(2)     TGGACCACACATTCT ACTGTGTCTAC  GGA GA GTCA CATACC

C3(1)     GAAGTGTCTGATGTACTAACTGGAGGATC
```

1ST INTRON/2ND EXON BOUNDARY

```
C3(1)     AATGTAATGATAGGTAAGCACTAATGAGTTCAAACCTCCATTCATGAATGTCATGTCTTT
C3(2)     AATGTAATGATAGGTAAGCACTAATGAGTTCAAACCTCCATTCATGAATGTCATGTCTTT

                                                    150       *
C3(1)     CTAGGTTCTGGCTGCAGTATTCTAGATGAAGTTATTAGAGGTACAATTAACTCAACTGTG
C3(2)     CTAGGTTCTGGCTGCAGTATTCTAGATGAAGTTATTAGAGGTACAATTAATTCAACTGTG
```

2ND EXON/2ND INTRON AND 2ND INTRON/3RD EXON BOUNDARIES

```
                      300
                       *                                  *  *    *       *
C3(1)     GTTGGCGTGATGATGGTAACTTTGAATTTGTTTTGTTTTGTT GTTTT TTGTTTTTTGT
C3(2)     GTTGGCGTGATGACGGTAACTTTGAATTTGTTTTGTTTT TTTGTTTTGTTGTTTTT GT

C3(1)     TTTGTAAACTGTGAAATCACTAACAAGGCGGAAGGAGAGAACTAGCTTCCAACAGACAGT
C3(2)     TTTGTAAACTGTGAAATCACTAACAAGGCGGAAGGAGAGAACTAGCTTCCAACAGACAGT

                                 *   **              * *   **  * *   *
C3(1)     TGTTCTCTGATTTCCCACACTCATGAGGCCACCACACAGAA ATACACATG TTAATCAT
C3(2)     TGTTCTCTGATTTCCCACACTTATG  GCCACCACACAGAATACACA  TGCT AATTAT

C3(1)     ACACATTCTAGA
C3(2)     ACACATTCTAGAGGTCAATGCCAGCAGCAAACCACCGAACTGAGAACAGGACCCCTTGGG

C3(2)     GGAAACTAGATGAAGTATTGAAAGAGTTGAAGGAGCTTGCAACCCCATAAAAACAACGAT

C3(2)     ACCAACCCACCAGAGCTTCCAGGGACTAAACCACTACCGGAAGACTATACATGGACTGAC

C3(2)     CCAGGGCTCCCAAATGCATAGGTAGCAGAGAATAGCTTTGTTGGGGTACCAGTGAAAGGG

C3(2)     AAAAGCTTTGGACTTGCCAAGGTTGGACCCCCAGTGCAGGAAAACATGGGGGGAGGGCAA

C3(2)     TGAGGGGGATGTATAGGGAGACACCGTATGGGAGAGAGGGAGGGGAAGGAATGGGGGCTT

C3(1)                                    CCTTTAAATGTAAGTAAGGAAATATATAATAA
C3(2)     ATGGACAGGAAACCAGGAAGGGGAATAACCTTTAAATGTAAGTAAGGAAATATATAATAA

                                                          *
C3(1)     AAAAATTACTAAGAAAAATGAACAAAAATGCTAGCAAGAAAAAAT GGTTTTTATCCTGT
C3(2)     AAAAATTACTAAGAAAAATGAACAAAAATGCTAGCAAGAAAAAATTGGTTTTTATCCTGT

                      *                                   *
C3(1)     GACTTTTAAGGACAATTCTAAATTTAAGACATCTCTTTGTTTTG TTCTCCAGGAGGCAA
C3(2)     GACTTTTAAGGACA TTCTAAATTTAAGACATCTCTTTGTTTTGGTTCTCCAGGAGGCAA

                     325
C3(1)     TATTTAACAGTGAAAGCTGTCAACAGC
C3(2)     TATTTAACAGTGAAAGCTGTCAACAGC
```

2.4 Discussion

The most important finding from the DNA sequencing of the genes for C3 is that only C3(1) can encode the secreted C3 polypeptide (Peeters et al., 1981) and the C3 cDNA clone, pA34 (Parker et al., 1983). From Fig. 10 it can be seen that the C3(2) gene possesses several important sequence differences, particularly within the second exon where there are eight nucleotide changes compared to C3(1), as illustrated in Fig. 12. One of these substitutions is silent while the others combine to produce six amino acid differences relative to the C3 polypeptide sequence (Peeters et al., 1981). There are also three substitutions and one insertion within the 3' non-translated region of C3(2) compared to the C3(1) sequence.

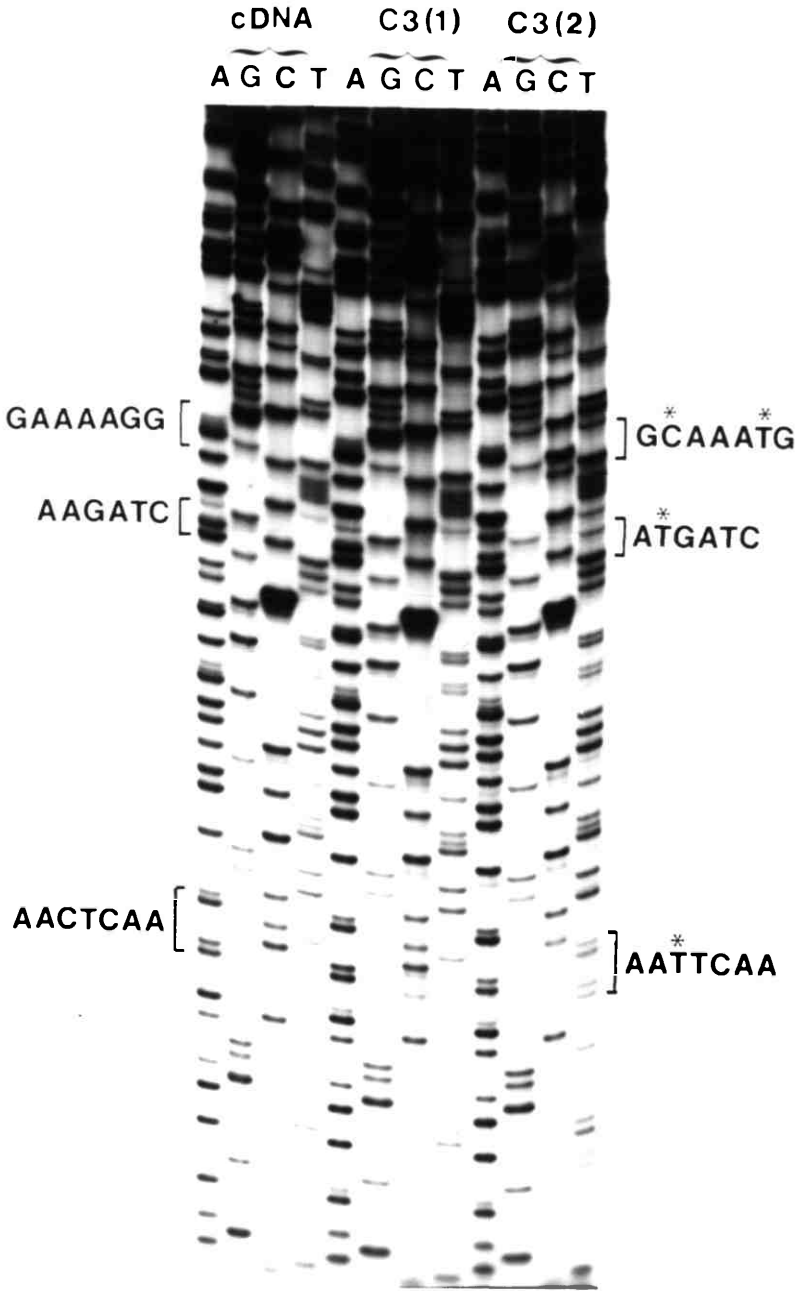
Although these sequence differences within C3(2) are crucial, the similarity between the two genes is most marked and C3(2) is nevertheless 97.7% homologous to pA34. Both C3 genes possess two introns of 1.7 kbp and 0.69 kbp which interrupt the coding regions at the same points to give exons of 112 bp, 188 bp and 214 bp; close to the sizes predicted from R-loop analysis. Furthermore, the first exon, which encodes the signal peptide removed during secretion of C3, is identical in sequence in C3(1) and C3(2). The overall distribution of differences between the coding portions of the two genes probably accounts for certain results from the Southern blotting and R-loop experiments. Hybridization to the C3(2) middle exon was observed to be reduced or absent respectively, which may be explained by the majority of the substitutions being within this exon. Conditions for DNA-RNA hybridization are likely to be much more critical than those for DNA-DNA hybridizations. Consequently, eight base changes in 188 (4.25%) was sufficient to prevent the observation of an RNA/DNA hybrid during R-loop analysis, but only reduced the degree of DNA hybridization in the blotting experiments.

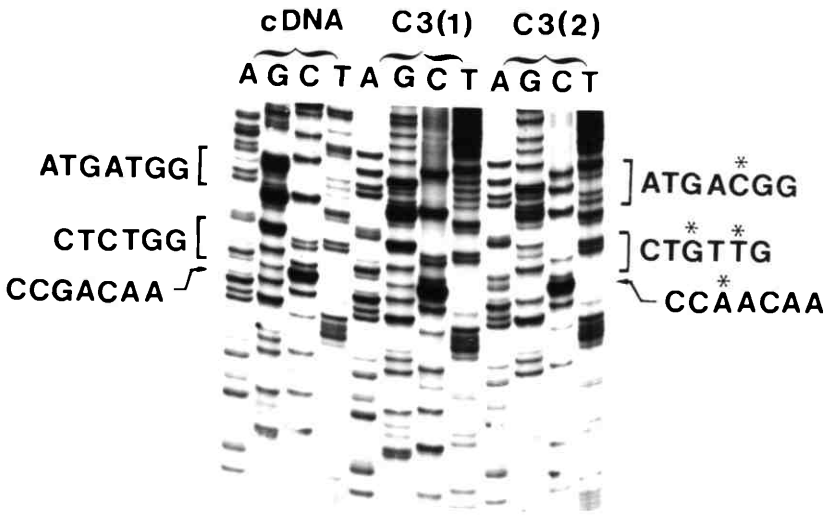
A comparison of the flanking DNA also shows the two C3 genes to be very similar. A total of 585 bp of 5' flanking DNA is 96.6% homologous, while 150 bp of 3' flanking DNA is 97.3% homologous. A true assessment of the similarity between the corresponding intervening sequences of both genes is not really possible with the present fragmented data, but the sequences shown in Figure 11 are ~95% homologous overall.

The sequences flanking the introns are identical in corresponding positions in the two C3 genes and agree well with published consensus

Figure 12. Second exon differences between C3(1) and C3(2).

The 900 bp Pst I/Bgl II restriction fragment, containing second exon and intron sequences, was purified from both λ6 and λ8 and cloned into mp 8, digested with Pst I and Bam HI. The corresponding fragment of 250 bp (containing no intervening sequence) was similarly cloned from the C3 cDNA clone, pA34. M13 clones containing each of these sequences were 'dideoxysequenced' and run on a 6% acrylamide gel for two hours to display the eight nucleotide differences (indicated by asterisks, *) between C3(2) on the one hand and pA34 and C3(1) on the other. The upper figure shows the whole gel, while the lower is a shorter exposure of the top of the gel alone. As the coding strand was cloned in each case the sequence read off the gel (5' ⟶ 3' = bottom ⟶ top) corresponds to the non-coding strand and may be compared to Fig. 10.


cDNA
C3(1)
C3(2)
A G C T A G C T A G C T
GAAAAGG
AAGATC
AACTCAA
GCAAATG
ATGATC
AATTCAA


cDNA
C3(1)
C3(2)
A G C T A G C T A G C T
ATGATGG
CTCTGG
CCGACAA
ATGACGG
CTGTTG
CCAACAA

sequences for splice sites (Breathnach et al., 1978; Mount, 1982; Fig. 10). Three possible classes of introns have been described: class 0 interrupt the reading phase between codons; class 1 interrupt between the first and second base of a codon; class 2 interrupt between the second and third base of a codon (Sharp, 1981). Using this classification, the introns in the C3 genes belong to classes 0 and 1 respectively. Therefore, the codon for the amino acid shown at the beginning of the second exon in Fig. 10 must be created by the splicing of the primary transcript.

To see if the sequences determined for the two C3 genes contained homologies to any other published sequences, a computer search comparing C3 with the E.M.B.O. sequence library was performed. This lead to the identification of an element present in the 5' flanking DNA of both genes (at -450 to -370, Fig. 10) that resembles the Alu-like moderately repetitive elements found dispersed in mammalian DNA (reviewed Jelinek and Schmid, 1982). This element is identical in sequence in C3(1) and C3(2) and is flanked by AG-rich sequences. A fuller study and description of this repetitive element is presented in the final results section (R5).

R3: MAPPING OF C3 TRANSCRIPTS TO C3(1)

3.1 Introduction

This section describes experiments intended to delineate those parts of the C3(1) gene that are transcribed into C3 RNA. Having sequenced a gene for which a cDNA and protein sequence had already been determined, it was known that the gene was split and also roughly where the 5' end of the mRNA might map. However, cDNA clones are rarely full-length at the 5' end and sometimes lack 3' sequences, for example pA34 cannot be full-length at its 3' end as it does not include a polyA tail (Parker et al., 1983).

To deduce the transcription start site(s), prostatic total $p(A^+)$ RNA (~10% comprising C3 mRNA) was hybridized to 5' end-labelled cloned genomic fragments (Weaver and Weissman, 1979) in primer extension and S1 mapping procedures. In the first technique a small primer, whose unlabelled end falls short of the 5' end of complementary mRNA, is extended using reverse transcriptase to give a blunt-ended hybrid. S1 mapping is the converse of this process as the unlabelled end of the DNA probe extends beyond the 5' end of the mRNA and is cut back using S1 nuclease so that again the ends of the hybrid abut. The products of both experiments are run out on sequencing gels and autoradiographed so that the observed change in length of the labelled probe or primer may be used to map C3 transcripts to the genomic sequence. The 3' ends of transcripts may be similarly mapped using 3' end-labelled S1 probes.

These same techniques can also be applied to check all the exon/intron boundary assignments deduced by comparison of C3 genomic and cDNA sequences. However, while many genes exhibit a duplication of sequences at their intron boundaries, making the positioning of splice points ambiguous (Sharp, 1981), this is not true for the C3 genes. Consequently it was felt that the intron/exon boundaries had been accurately determined by sequence comparison alone.

3.2 Mapping 5' Termini of C3 mRNA

Primer extension was generally performed using a 68 b primer prepared from the cDNA clone pA34 as illustrated in Fig. 13A. An additional 26 b primer labelled at the Bst EII site in pA34 was also used as in Fig. 13C. The S1 probe was prepared from the C3(1) subclone

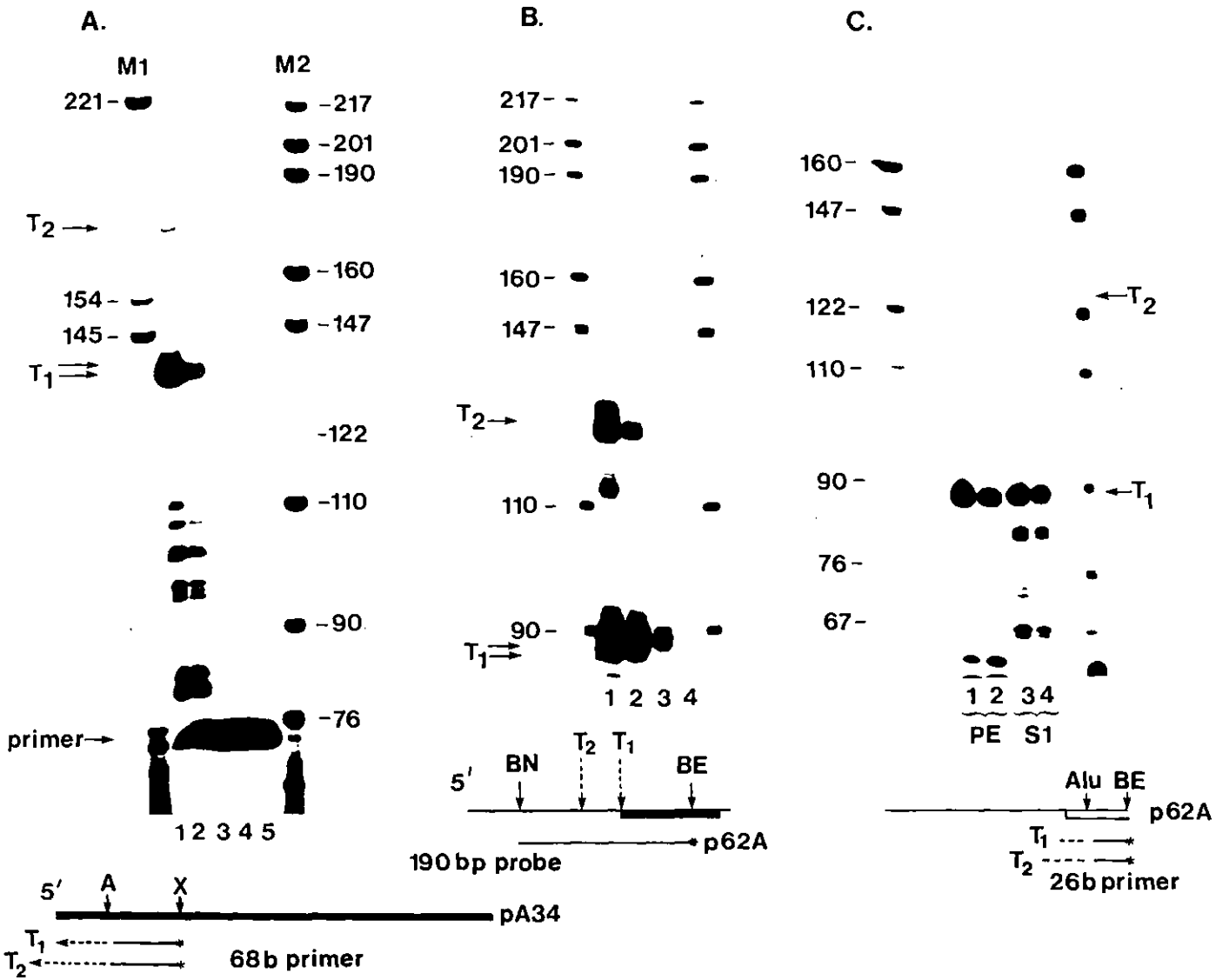
Figure 13. Mapping 5' ends of prostatic C3 transcripts.

(A) Primer extension mapping was carried out using a $[^{32}P]$-labelled 68 b primer generated from the cDNA clone, pA34. The clone (5 µg) was digested with Xba I (X), 5' end-labelled and redigested with Alu I (A) before strand separation. All the samples contained 30,000 c.p.m. primer and 20 µg RNA, comprising yeast RNA plus ventral prostate total $p(A^+)$ RNA as follows: track 1, 1.0 µg; track 2, 0.1 µg; track 3, 0.01 µg; track 4, 0.001 µg; track 5, control.

(B) S1 mapping was carried out using a 190 bp, $[^{32}P]$-labelled probe generated from p62A, a subclone of the first exon of C3(1). The clone was digested with Bst E_{II} (BE), 5' end-labelled and redigested with Bst NI (BN) before gel purification. All samples contained 50,000 c.p.m. probe and 25 µg RNA comprising yeast RNA plus prostate $p(A^+)$ RNA as follows: track 1, 1.0 µg; track 2, 0.1 µg; track 3, 0.01 µg; track 4, control.

(C) S1 mapping and primer extension were directly compared using the 190 bp probe from B above and a shorter primer (26 b) prepared from pA34. The clone was digested and labelled at the Bst E_{II} (BE) site corresponding to that used to label the 190 bp probe. The second digest again used Alu I (A), as for the 68 b primer in A above. Tracks 1-3 contained 30,000 c.p.m. 26 b primer while tracks 4-6 contained 50,000 c.p.m. 190 bp probe. All samples contained 25 µg RNA comprising yeast RNA plus ventral prostate $p(A^+)$ RNA as follows: tracks 2 and 4, 1.0 µg; tracks 3 and 5, 0.5 µg; tracks 1 and 6, control.

In all cases the markers are 5' end-labelled Hpa II digested pAT 153, except M1 in A which are 5' end-labelled Hinf I digested pAT 153.


A.
M1
M2
221
217
201
190
160
147
122
110
90
76
T2
T1
154
145
primer
1 2 3 4 5
5'
A
X
pA34
T1
T2
68b primer
B.
217
201
190
160
147
T2
110
90
T1
1 2 3 4
5'
BN
T2
T1
BE
p62A
190bp probe
C.
160
147
122
110
90
76
67
T2
T1
1 2 3 4
PE
S1
Alu
BE
p62A
T1
T2
26b primer

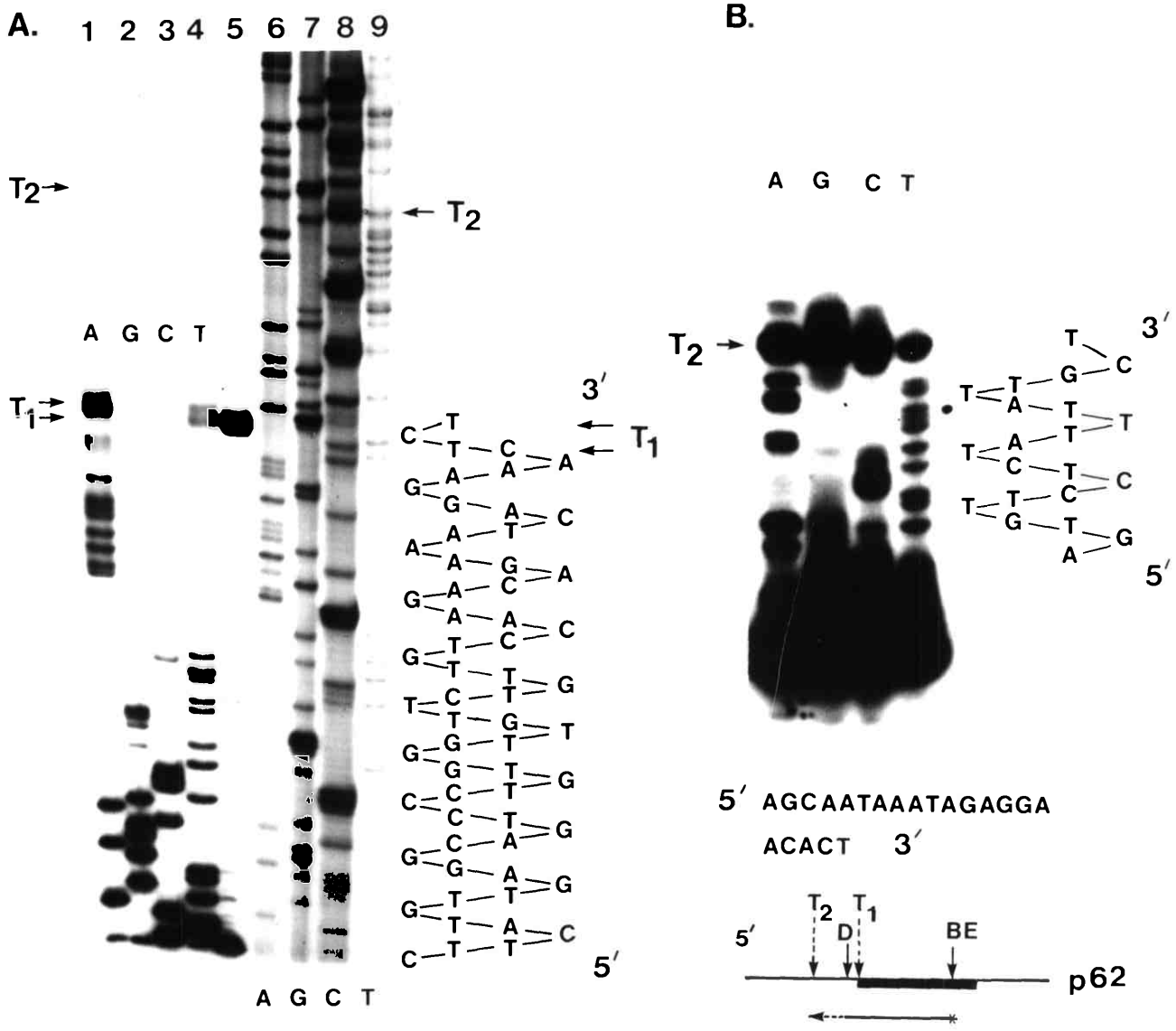
p62 and was 190 bp long as shown in Fig. 13B, but a longer probe of ~850 bp has also been used (see Fig. 19B). Each of these labelled DNA fragments was hybridized to a range of concentrations of ventral prostate p(A^+) RNA and the products after enzyme modification were run out on 6% sequencing gels as shown in Fig. 13. The intensity of the bands across the tracks is a roughly linear relation indicating that the DNA was in excess and that all possible transcripts were hybridizing and being detected within the sensitivity of the techniques used. Both S1 mapping and primer extension show clearly that there are two C3 transcription initiation regions designated T_1 (major) and T_2 (minor). That these two regions coincide for both techniques can be ascertained using the markers and known initial lengths of the DNA fragments, but has been confirmed by using the 26 b primer and 190 bp probe in parallel, which resulted in the products from both experiments running at equivalent positions on the gel (Fig. 13C). The long (850 bp) S1 probe was used to assay for any transcripts initiating further upstream (see Fig. 19B, track 1), but T_1 and T_2 appear to be the only initiation regions used in vivo.

In mapping T_1 and T_2 to the genomic sequence minor artefacts, present in both techniques, have been considered. The 'ladders' produced in primer extension reactions are unlikely to represent additional transcription starts as they were not present in S1 experiments. Instead, they are probably due to reverse transcriptase pausing at particular sequences (rather than mRNA breakdown) as they form a distinct and constant pattern. Digestion with S1 can produce artefacts owing to enzyme 'nibbling' at blunt-ended RNA/DNA hybrids, making each gel band into a doublet or triplet, but addition of too little enzyme causes incomplete digestion of hybrids (e.g. Fig. 13B, track 1). In view of these ambiguities, the transcript starts were examined more closely by primer extension under chain terminating conditions. Using the 68 b primer, the resulting sequence of the 5' end of transcripts initiating at T_1 was compared with the corresponding genomic sequence, as shown in Fig. 14A. It is apparent that transcripts from T_1 have heterogenous 5' ends that map within the sequence 5' CCAGAG 3'. Although the mRNA sequence was a little unclear at the 5' end (possibly caused by inhibition of reverse transcriptase by methylated ribonucleotides at the cap site) T_1 RNA's probably originate equally from either A residue within this sequence. It is worth noting here that the second A also coincides with the 5' end of the C3 cDNA clone,

Figure 14. Primer extension sequence analysis of C3 transcripts.

(A) The 5' end sequence of transcripts initiating at T_1 was determined by primer extension analysis under chain termination conditions using the $[^{32}P]$-labelled 68 b primer described in Fig. 13A. Tracks 1-4 each represent 10,000 c.p.m primer hybridized to 2.0 µg total ventral prostate p(A^+) RNA. Track 5 shows primer extension under normal conditions using 30,000 c.p.m. primer hybridized to 0.25 µg p(A^+) RNA. All samples contained carrier yeast RNA to 20 µg. Tracks 6-9 show the sequence of the genomic coding strand for comparison. The 2.1 kb Pst fragment of p11B1 (see Appendix) was cloned into M13 mp 7 and the non-coding strand clone sequenced using the 68 b primer.

(B) The 5' end sequence of transcripts initiating at T_2 was determined using a $[^{32}P]$-labelled 95 b primer prepared from p62A by 5' end-labelling at the Bst E_{II} (BE) site and redigesting with Dde I (D) before strand separation. Each track represents 20,000 c.p.m primer hybridized to 12.5 µg prostate p(A^+) RNA which was primer extended under chain terminating conditions. The sequence produced, representing the coding strand, is shown adjacent to the gel, with the complementary non-coding sequence written underneath. A dot next to the middle band of a triplet within the T-track marks the nucleotide that is mutated in C3(2).


A.
1 2 3 4 5 6 7 8 9
T2
T2
A G C T
T1
3′
T1
5′
A G C T
B.
A G C T
T2
3′
5′
5′ AGCAATAAATAGAGGA
ACACT 3′
T2 T1
5′
D
BE
p62

pA34, hence has been designated nucleotide +1 of the transcribed gene (Fig. 10).

The transcription initiation region T_2 appears to map 36/38 b upstream of T_1 to an A residue at -37. Unfortunately, there are several artefact bands in the primer extension sequencing analysis in Fig. 14B, so this experiment was not wholly conclusive in terms of precisely assigning the 5' end of transcripts originating at T_2. However, the results are interesting as it can be seen that the 5' end of these transcripts 5' AGCAATAAATA...3' is equivalent to that encoded by C3(1) and that they do not contain the mutated sequence 5' AGCAATAGATA...3' which is encoded at this point by C3(2) (Fig. 10).

3.3 Mapping 3' Termini of C3 mRNA

As shown in Fig. 15, two 3' end-labelled S1 DNA probes have been used to map the 3' ends of C3 transcripts. Using the 312 bp probe A, hybridized to ventral prostate $p(A^+)$ or total RNA, a single protected band of 165 b was produced (Fig. 15A), implying C3 transcripts end within the sequence 5' CAA...3' (Fig. 10). With the longer probe B and a long exposure (1 week) of the autoradiograph, additional, faint bands were observed indicated by (?) in Fig. 15B. These would map some 200-400 bp further downstream of the main band, which presumably represents the (major) mature 3' terminus of C3 mRNA.

3.4 Effect of Testosterone on Transcription from T_1 and T_2

The steady-state ratio of C3 transcripts initiated from T_1 and T_2 was determined by densiometric scanning of autoradiographs from several primer extension and S1 5' end mapping experiments. Examples of such scans are shown in Fig. 16. The ratio of T_1 to T_2 was found to be 18:1, averaged from 15 separate scans covering three different preparations of ventral prostate $p(A^+)$ RNA.

To see if this ratio altered with the androgen status of the tissue, primer extension reactions were performed using the 68 b primer and prostate nuclear RNA from hormonally manipulated rats (Fig. 17). The resulting gel autoradiograph was scanned across all three tracks and showed that both transcripts fell to 20% of normal levels in castrated animals, but treatment with testosterone restored them to 60% of normal levels. Hence, the relative ratios of the two families of C3 transcipts do not alter during hormonal manipulation.

Figure 15. Mapping 3' ends of prostatic C3 transcripts.

S1 mapping of 3' ends was carried out using probes generated from p61 by 3' end labelling at the Bgl II (Bg) site and redigesting either with Alu I (A) to give a 312 bp probe A, or using Pst I (P) to give a 600 bp probe B.

(A) All samples contained 30,000 c.p.m. probe A plus 20 µg RNA comprising yeast RNA with prostate RNA as follows: track 1, control; track 2, 0.01 µg p(A^+) RNA; track 3, 1.0 µg p(A^+) RNA; track 4, 10 µg total RNA; track 5, 20 µg total RNA. The gel was run for 3 hrs, P indicates the position of reannealed probe. The markers were 5' end-labelled Hpa II digested pAT 153.

(B) Both samples contained 50,000 c.p.m. probe B together with RNA as follows: track 1, 25 µg total prostate RNA; track 2, 25 µg yeast carrier RNA. The gel was run for 3 hrs, P indicates the position of reannealed probe and (?) show possible minor 3' ends or C3 transcription termination sites. The markers were 5' end-labelled Hinf I digested pAT 153.

In both A and B the band corresponding to the (major) mature 3' terminus of C3 p(A^+) RNA is labelled 3' end.

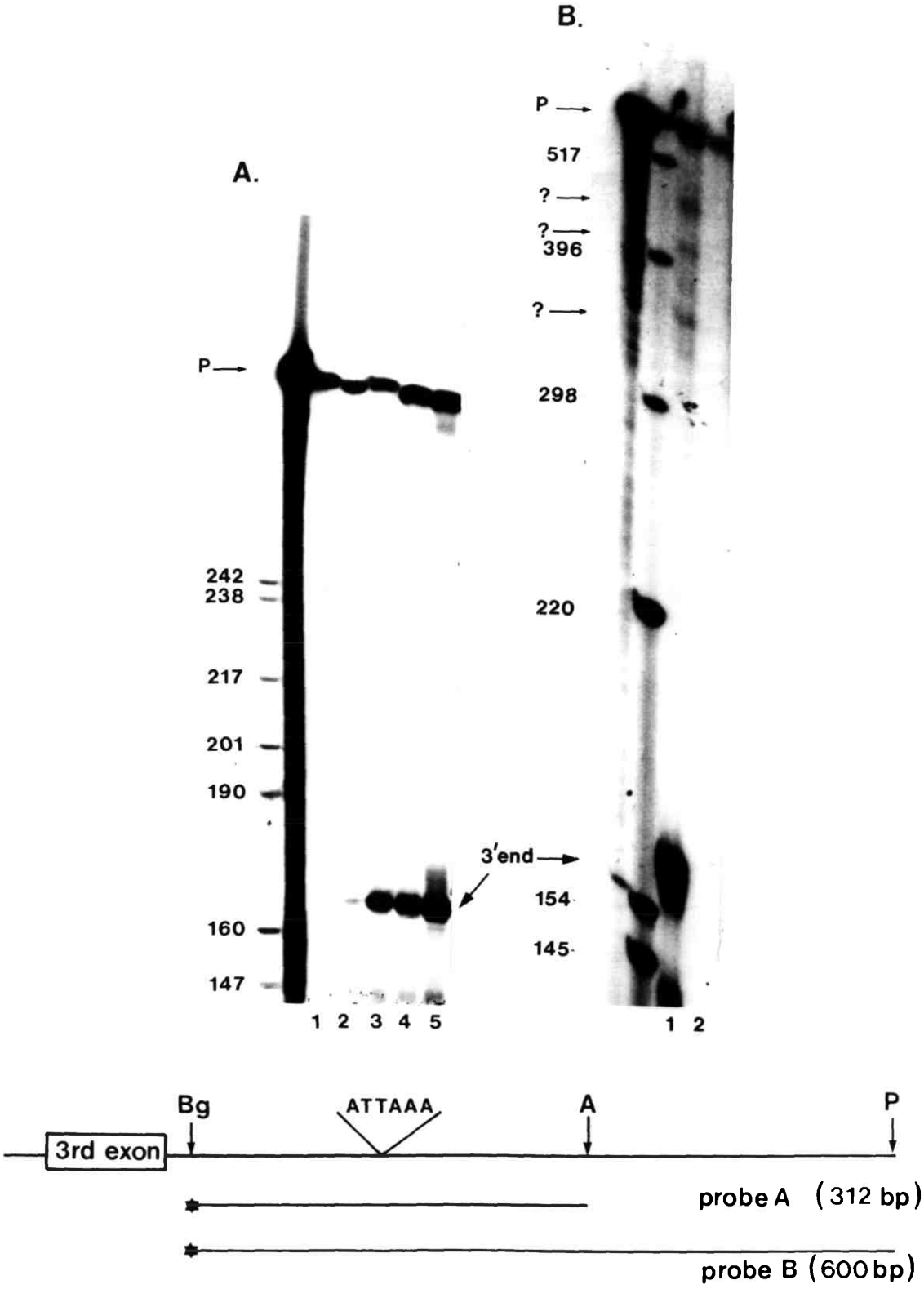
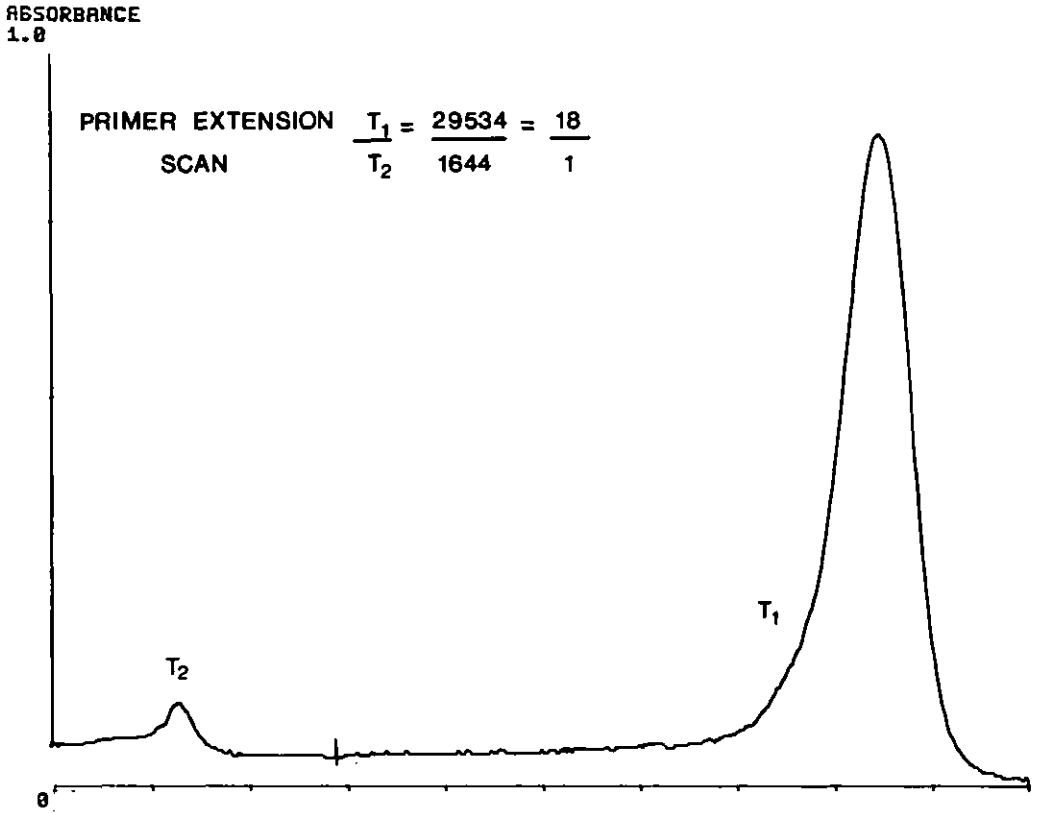

A.
B.
P
242
238
217
201
190
160
147
3'end
1 2 3 4 5
P
517
?
?
396
?
298
220
154
145
1 2
Bg
ATTAAA
A
P
3rd exon
probe A (312 bp)
probe B (600 bp)

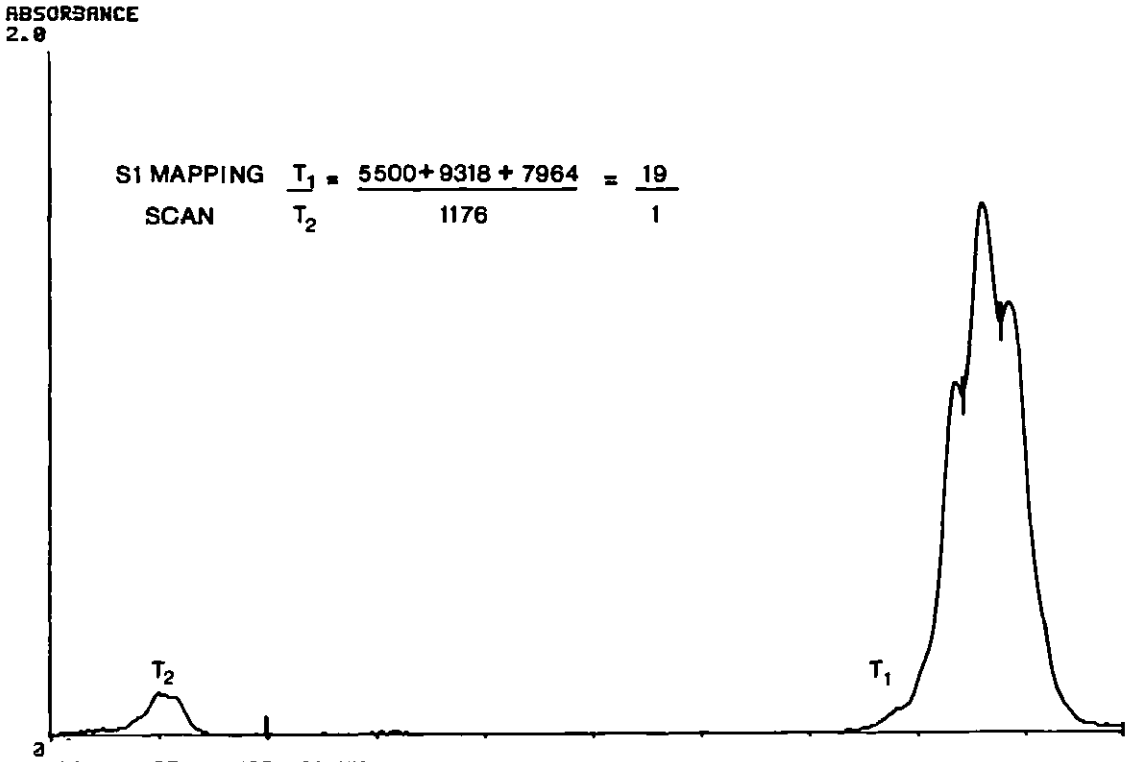
Figure 16. Densiometric scans of S1 and primer extension 5' mapping gel autoradiographs.

One example of a gel scan using each mapping technique is shown, both possessing two peaks representing the mapping of mRNA 5' ends to T_1 and T_2. In S1 mapping gel autoradiographs the T_1 peak was always seen as at least a triplet, possibly due to 'nibbling' of S1 (see text). The relative intensities of the T_1 and T_2 bands were determined from the ratio of the integrated areas under each peak.



LENGTH OF X-AXIS= 60 MM.
TOTAL INTEGRAL VALUE= 31178

PEAK	POSITION	REL%	INTEGRAL	PEAK	POSITION	REL%	INTEGRAL
1	7.9	5.27	1644	2	50.8	94.72	29534

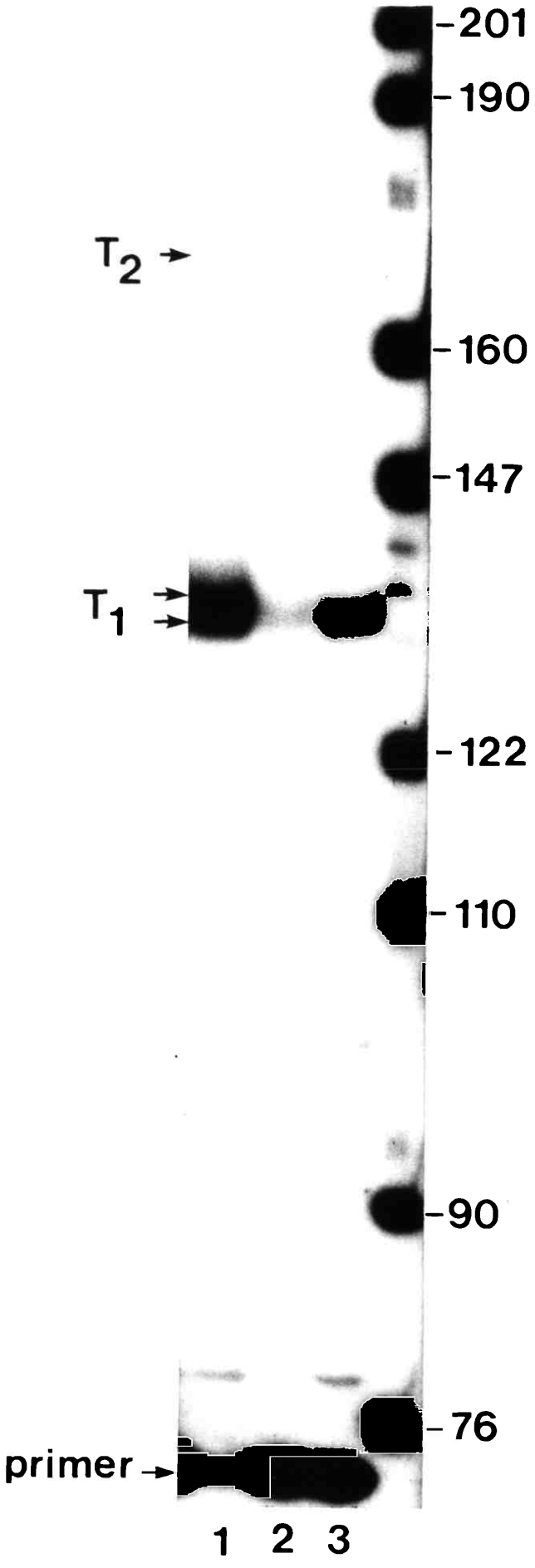


LENGTH OF X-AXIS= 80 MM.
TOTAL INTEGRAL VALUE= 23958

PEAK	POSITION	REL%	INTEGRAL	PEAK	POSITION	REL%	INTEGRAL
1	8.2	4.9	1176	2	67.0	22.95	5500
3	69.8	38.89	9318	4	71.8	33.24	7964

Figure 17. Effect of androgens on C3 transcript levels.

Primer extension was carried out using the [^{32}P]-labelled 68 b primer described in Fig. 13A. Each sample contained 35,000 c.p.m. primer with 20 µg carrier yeast RNA and 2.5 µg rat ventral prostate nuclear RNA derived as follows: track 1, normal animals; track 2, animals 3 days after castration; track 3, 3-day castrates after 24 hrs with testosterone. Nuclear RNA samples were provided by M. Page. The markers were [^{32}P]-labelled Hpa II digested pAT 153.


201
190
T2
160
147
T1
122
110
90
76
primer
1 2 3

3.5 Discussion

Mapping the 5' and 3' ends of C3 transcripts has shown that there are two distinct families which share a unique mature 3' terminus, but have 5' termini that map ~36 b apart. The two types of transcript are present at a constant ratio of 18:1 in ventral prostate and it is apparent that androgens have a co-ordinate effect on steady state levels of both transcripts.

Knowing where the transcriptional initiation and termination sites map on the gene allows the flanking DNA sequences to be examined for putative transcriptional control regions (see Fig. 10) as discussed below.

Upstream of T_1 at -35 in C3(1) the sequence 5' CAATAAATA 3' occurs, which closely resembles in both sequence and position relative to the cap site the 'TATA' box thought to be (part of) the RNA polymerase II promoter. In C3(2), however, the sequence is mutated to 5' CAATAGATA 3' and it is therefore interesting that this A→G transversion is the only base change between the two genes from nucleotide -235 to within the first intron (see Figs. 10 and 11).

The transcriptional start site T_2 maps 33 b downstream of the sequence 5' TATTTAT 3' present in both C3(1) and C3(2). While not possessing obvious homology to a 'TATA' box, this sequence is not without precedent as it also occurs in a corresponding position in front of the SV40 early (Benoist and Chambon, 1981) and adenovirus 5 E1a (Maat *et al.*, 1980) genes.

The 3' end of C3 mRNA maps 12 bp downstream of a sequence 5' ATTAAA 3' found in the non-coding DNA of both C3 genes. This closely resembles the putative polyadenylation signal observed in most other eucaryotic protein-coding genes (Proudfoot and Brownlee, 1976). Exact homologues of the C3 sequence have been found in the 3'non-translated DNA of, for example, the chicken lysozyme (Jung *et al.*, 1980) and mouse pancreas α-amylase (Hagenbüchle *et al.*, 1980) genes. Other sequences have also been noted to occur frequently around gene polyadenylation sites (Benoist *et al.*, 1980) although their functional significance has not been assayed. For example, many mature transcripts, including those of C3, end CA(A...), leading to an ambiguity as to the exact 3' terminus of the gene coding region. The C3 genes also show good homology to a consensus sequence found in several genes between the polyadenylation signal and the actual site of polyadenylation

(Benoist et al., 1980):-

Consensus	5'	T	T	T	T	C	A	C	T	G	C	3'
		|	|		|	|		|	|	|	|	
C3(1 and 2)	5'	T	T	A	T	C	T	C	T	G	C	3'

However, as described in the Introduction, the site of polyadenylation is not believed to coincide with actual transcriptional termination. True termination sites for C3 transcripts have not been rigorously searched for, but the evidence presented in Fig. 15B indicates that there may be three possible termination sites 200-400 bp further downstream of the polyadenylation site. Alternatively, the 3' ends mapped in Fig. 15B may represent minor forms of mature C3 mRNA.

Clearly though, the mapping of C3 transcripts to C3(1) has shown that the DNA flanking the gene contains control sequences whose presence is now considered mandatory within protein coding genes.

R4: IS C3(2) EXPRESSED *IN VIVO*?

4.1 Introduction

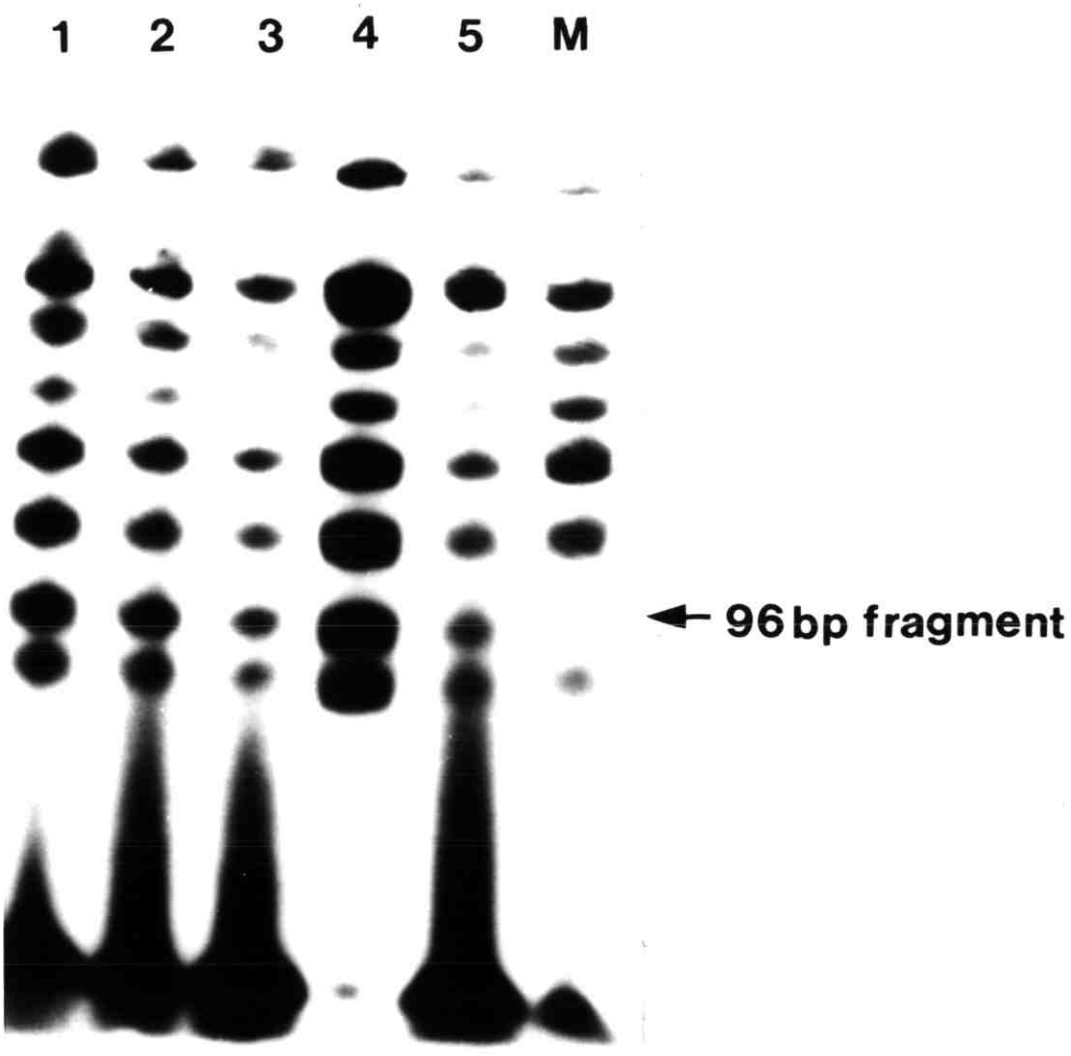
In the previous section the mapping of C3 transcripts was described in terms of C3(1) alone being expressed *in vivo*. Certainly C3(2) is probably not highly active in the ventral prostate because the sequence information shows that it cannot encode the C3 polypeptide. Furthermore, all of the C3 cDNA clones isolated from rat prostate $p(A^+)$ RNA were clearly derived from C3(1) transcripts (see Fig. 18). Nevertheless, the C3(2) coding sequence does not reveal any obvious aberrations, compared to C3(1), that would render it a pseudogene incapable of producing viable transcripts. For example, C3(2) does not possess internal stop codons or frame shift mutations within the coding region and bears exactly the same putative splice and polyadenylation signal sequences as C3(1). Moreover, at the 5' end the two genes show only one nucleotide difference from -235 to within the first intron. This one substitution at -30 does fall within the T_1 'TATA' box and may cause reduced expression, but is unlikely alone to result in complete failure of the C3(2) gene to be transcribed, as expression could still occur from T_2. However, a similar A→G transversion at -28 in the β-globin gene of a β-thalassemia patient is believed to account for the 10-fold reduction in β-globin transcription observed in this individual (Orkin *et al.*, 1983). The resulting β-globin 'TATA' box 5' CATAGAA 3', is reminiscent of that found for C3(2) T_1 thus casting a doubt on the transcriptional viability of this gene.

That C3(2) is capable of producing transcripts initiating at T_1 and T_2 in an androgen regulatable manner was shown by stable transfection of the cloned gene into mouse S115 cells (Page and Parker, 1983). C3(1) was also active in this assay. However, the relative transcriptional efficiencies of the two genes could not be compared due to the difficulty in determining the exact number of expressing copies of the transfected gene present in each cell line.

For this reason, it was attempted to directly compare transcription from C3(1) and C3(2) by using restriction fragments, corresponding to the 5' end of the genes, as templates in an *in vitro* transcription assay (Manley *et al.*, 1980). The study was then extended by trying to directly assay for a population of C3(2) transcripts in normal prostate $p(A^+)$ RNA.

Figure 18. Analysis of the five available C3 cDNA clones.

The C3 cDNA clones were assayed by restriction enzyme analysis to determine if any were derived from a transcript of C3(2). The C3(1) cDNA clone, pA34, has two internal Hinf I sites, at nucleotide +273 and +369. However, at +277 C3(2) shows a base substitution compared to C3(1), changing the Hinf I recognition sequence 5' G A C T C 3' to 5' G A G T G 3', which is not a substrate of the enzyme. Therefore the 96 bp Hinf I restriction fragment within pA34 would be absent from any C3(2) cDNA clones. DNA (0.5 µg) from each clone was digested with Hinf I, 5' end-labelled and run on a small acrylamide gel. Direct autoradiography produced the pattern of bands shown for the five clones p5(1), pA10 (2), pA14 (3), pA34 (4), pA37 (5). Comparison with the 5' end-labelled Pst I/Hinf I digested pAT 153 markers revealed that all five clones contained an internal 96 bp Hinf I fragment and therefore were all derived from C3(1) transcripts.


1
2
3
4
5
M
96 bp fragment

4.2 In vitro transcription of the C3' genes

In the first experiment, the transcriptional activity of the C3(2) subclone p11B1 (5' end) was assayed in the HeLa cell free extract, in the presence of 0, 1 and 5 μg/ml α-amanitin. The RNA products were analysed by primer extension mapping of their 5' ends using the 68 b primer prepared from pA34 and the resulting gel autoradiograph is shown in Fig. 19A. Clearly C3(2) is capable of directing the synthesis of several products *in vitro*, including transcripts apparently initiating at T_1 and T_2. Although the accumulation of the majority of these products was not influenced by the addition of α-amanitin, levels of transcripts initiating at T_1 certainly declined with increasing concentration of the RNA polymerase II inhibitor.

With so many (artefactual) initiations a direct comparison of the two C3 genes in the *in vitro* assay was clearly going to be difficult. However, one experiment was attempted using analogous restriction fragments from each gene as templates in the HeLa extract and analysing the RNA products by S1 mapping (Fig. 19B). For both clones the restriction fragment assayed contained up to 800 bp of 5' flanking DNA and a pattern of many transcription initiations was observed covering the whole of this region. The majority of these initiations were observed for both genes, though a few unique gel bands did occur, presumably reflecting some sequence differences between the two genes at or prior to these start sites. Nevertheless, cloned fragments of C3(1) and C3(2) were apparently equally efficient at producing transcripts originating at T_1 and T_2 in this assay, though the level of such initiations was disappointingly low (Fig. 19B).

Given that C3(2) does not appear to be defective in producing correct initiation of C3 transcripts, a search was made for *in vivo* C3(2) mRNA.

4.3 Direct Probing for *in vivo* C3(2) mRNA

The strategy used in the search for C3(2) transcripts was based on the sequence differences between the two genes within their second exons. The scheme is shown diagrammatically in Fig. 20 and relies on S1 mapping, using the 3' end-labelled 260 bp Xba I-Alu I probe, being able to detect base differences between the probe prepared from cloned DNA of one gene and the RNA transcripts of the other. The sequence differences between C3(1) and C3(2) map so as to give expected protected

Figure 19. In vitro transcription of the C3 genes.

(A) The 2.1 kbp insert within the C3(2) subclone p11B1 (see Appendix) was separated from vector sequences by Pst I digestion and added to 100 µl in vitro transcription mixes to a concentration of 11 µg/ml of insert DNA. The RNA products were analysed by primer extension using the 68 b primer described in Fig. 13A. All the samples contained 50,000 c.p.m. primer with 20 µg RNA comprising yeast carrier RNA with: track 1, 1.0 µg prostate p(A^+) RNA; track 2, 0.1 µg prostate p(A^+) RNA; track 3, 0.01 µg prostate p(A^+) RNA; tracks 4-7, RNA from in vitro transcription with no added template (track 4), and with the C3(2) template transcribed in the presence of 0 (track 5), 1 (track 6) and 5 (track 7) µg/ml α amanitin. T_1 and T_2 are marked.

(B) Plasmids p61 and p81 were digested with Xba I generating a 2.6 kbp fragment from each clone corresponding to the 5' end of the gene. These digests were added to 200 µl transcription mixes to a concentration of 9 µg/ml of insert DNA. The RNA products were analysed by S1 mapping using an 850 bp probe prepared from p62A by 5' end-labelling at the Bst E_{II} site and redigesting with Eco RI. Each sample contained 40,000 c.p.m. probe and RNA as follows: track 1, 1.0 µg prostate p(A^+) RNA with 19 µg yeast carrier RNA; track 2, RNA from in vitro transcription of C3(1) (p61); track 3, RNA from in vitro transcription of C3(2) (p81). T_1 and T_2 are marked. The middle band in track 1 (--->) is probably an incomplete digestion product which is an artefact which occurs occasionally in S1 mapping experiments (see Fig. 13B, track 1).

In both A and B the markers were 5' end-labelled Hpa II cut pAT 153.

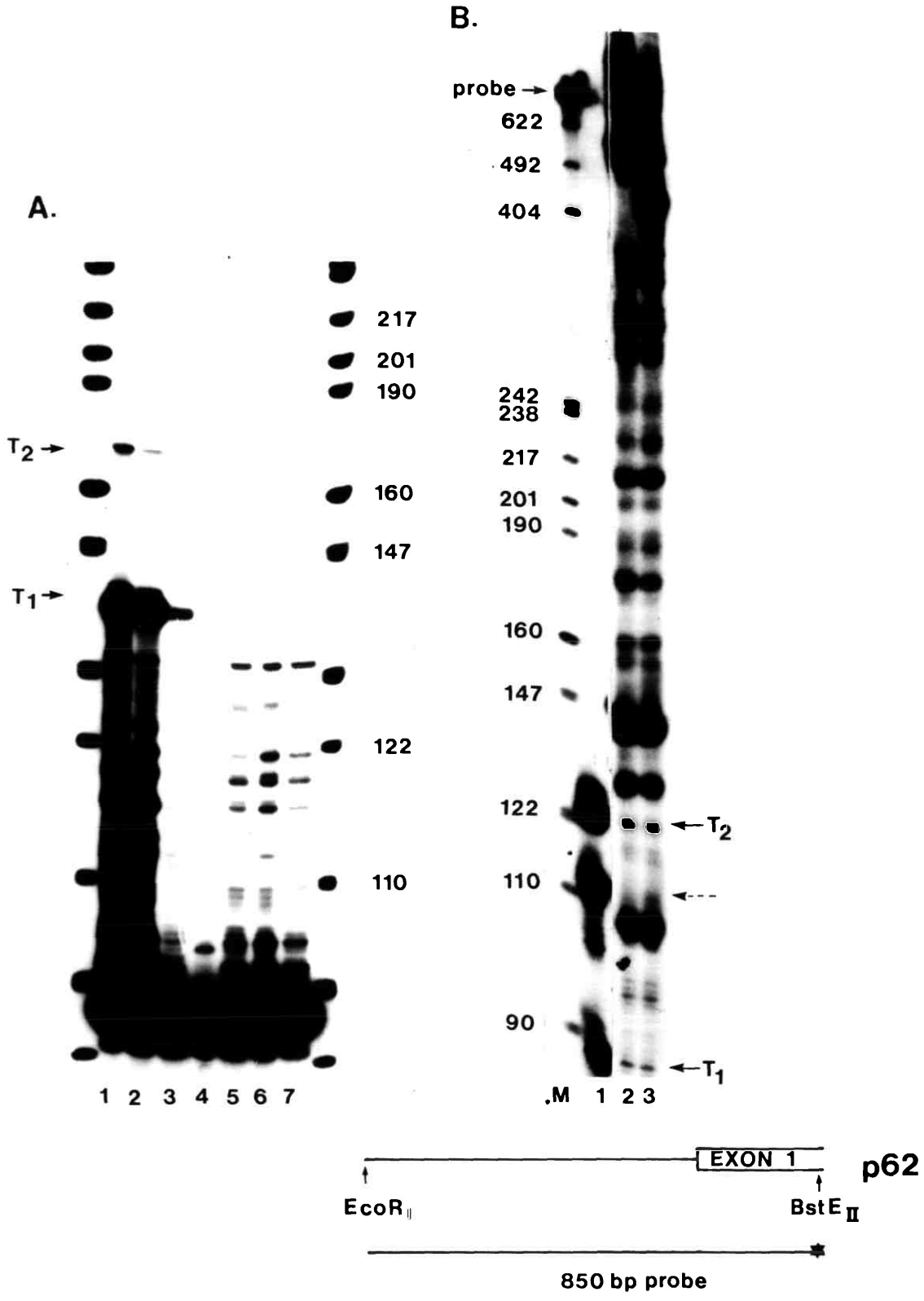

A.
T2
T1
217
201
190
160
147
122
110
1 2 3 4 5 6 7
B.
probe
622
492
404
242
238
217
201
190
160
147
122
110
90
T2
T1
.M 1 2 3
EXON 1
p62
EcoR II
BstE II
850 bp probe

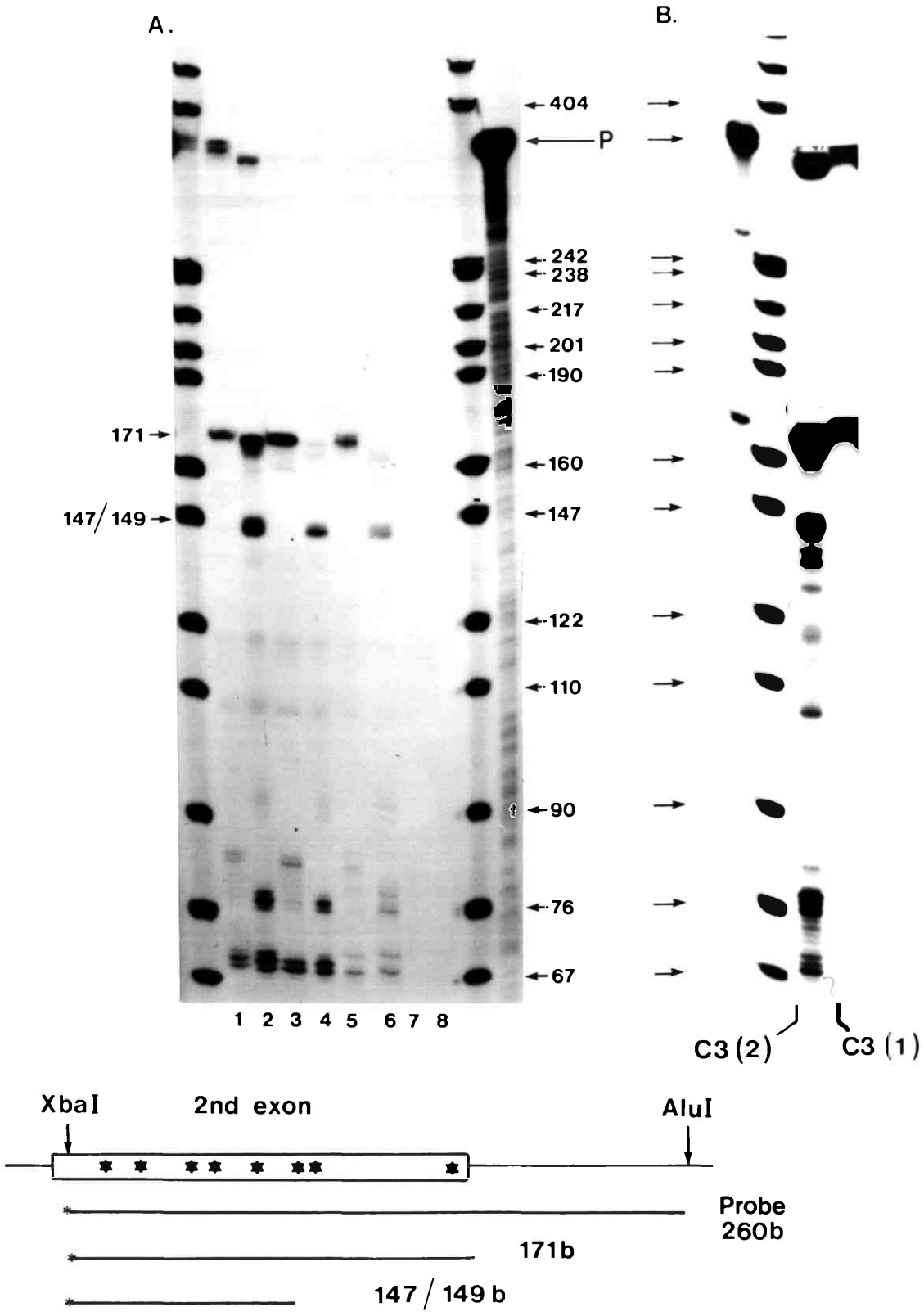
Figure 20. Direct probing for in vivo C3(2) mRNA.

The diagram illustrates the second exon of the C3 genes with asterisks indicating the points of mismatch between C3(1) and C3(2) within this region. A 3' end-labelled 260 bp probe was prepared from p61 and p81. Cloned DNA (10 µg) was digested with Xba I, resected further using exonuclease III and filled in with $[\alpha^{32}P]$-dCTP and -dTTP. The DNA was redigested with Alu I and the probe was gel purified.

(A) S1 mapping experiments were carried out with hybridization o/n at 58°C and various S1 digestion conditions. All the samples contained 2.5 µg prostate p(A^+) RNA, 20 µg yeast carrier RNA and 30,000 c.p.m. probe; C3(1) probe from p61 in tracks 1, 3, 5 and 7, C3(2) probe from p81 in tracks 2, 4, 6 and 8. S1 digestion conditions were as follows: tracks 1 and 2, 200 U/ml for 2 hrs at 37°C; tracks 3 and 4, 400 U/ml for 2 hrs at 37°C; tracks 5 and 6, 100 U/ml for 2 hrs at 45°C; tracks 7 and 8, 200 U/ml for 2 hrs at 45°C.

(B) Using S1 digestion conditions of 400 U/ml for 2 hrs at 37°C samples containing 10 µg p(A^+) RNA were assayed with 120,000 c.p.m. probe from C3(1) (p61) or C3(2) (p81).

P = reannealed probe. The markers were 5' end-labelled Hpa II digested pAT 153.


A.
B.
404
P
242
238
217
201
190
171
160
147/149
147
122
110
90
76
67
1 2 3 4 5 6 7 8
C3 (2)
C3 (1)
XbaI
2nd exon
AluI
Probe
260b
171b
147 / 149 b

probe sizes of 29 b, 80 b, 94 b, 138 b, 147 b, 149 b and 169 b. In practice, the standard conditions of RNA/DNA hybridization and S1 digestion are not stringent enough to detect these internal single base mismatches. Therefore, the Xba I-Alu I probe prepared from either gene would leave a protected fragment of 171 b, corresponding to the 3' end of the second exon. Consequently, the initial experiments using these probes were designed to find more exacting conditions of S1 mapping that would detect sequence differences between the C3(2) DNA probe and C3(1) mRNA and, hopefully, vice versa.

Using the C3(2) probe, it was shown that the hybridization temperature in formamide buffer could be raised from 52°C to 58°C o/n with little loss in intensity of the final protected band on the gel autoradiograph. Next a series of experiments were performed using this new hybridization temperature, but with varied conditions of S1 digestion, as shown in Fig. 20A. Although conditions were never found whereby all the base mismatches were detected by the S1 enzyme, under some conditions (Fig. 20A, tracks 4 and 6) all the C3(2) probe was cut at the two very close substitutions (at +277 and +279, Fig. 10) to leave protected fragments seen as a 147/149 b doublet. Under these same digestion conditions, however, a detectable 147/149 b doublet was not produced when using the C3(1) probe (Fig. 20, tracks 3 and 5) which would have indicated that some fraction of this probe had hybridized to a population of C3(2) transcripts within the ventral prostate.

Further experiments were attempted using the same conditions and proportionally increasing the amount of input p(A^+) RNA and probe, as shown in Fig. 20B where each track represents 10 µg p(A^+) RNA (that is ~1.0 µg C3 mRNA). Unfortunately, scaling the experiment up like this was not completely successful as not all of the C3(2) probe was 'chased' into the 147/149 b doublet. The remainder though left a protected band of 169 b, implying the base mismatch at the penultimate nucleotide of the exon was recognized. Again, however, there were no consistent smaller bands produced when using the C3(1) probe.

4.4 Discussion

Several genes have been examined as to their transcriptional viability, after the introduction of point mutations within the 'TATA' box, by assaying their activity in HeLa extracts. Generally, the introduction of a G residue into the 'TATA' box proved to be a down mutation as for the chicken conalbumin (Wasylyk *et al.*, 1980) and ovalbumin

(Zanucki-Schulz et al., 1982) genes. A comparison of C3(1) and C3(2) in vitro transcriptional template activity is really equivalent to such experiments, especially as a functional 'TATA' box has been generally shown to be absolutely required for authentic initiations in cell free systems (reviewed Breathnach and Chambon, 1981). Unfortunately the C3 genes are not good templates in cell free extracts and beyond showing that the 'promoters' of both genes appeared to be equally effective in the assay, these experiments were not pursued. It must be concluded though, from these results and the observations with C3(2) transfected S115 cells, that the base substitution within the C3(2) T_1 'TATA' box is of no functional significance and that there is no obvious reason why C3(2) should not be expressed in vivo.

Attempts to directly probe ventral prostate p(A^+) RNA for a population of C3(2) transcripts were largely frustrated by the difficulty in designing an experiment capable of distinguishing between two such similar sequences. Using S1 mapping and acrylamide gels some progress was made, but the sensitivity of this kind of approach is limited by the specific activity of the probe, which was only labelled at about 7 bases at its 3' end. In order to improve the chances of detecting C3(2) mRNA the products of further S1 mapping experiments were run on alkaline agarose gels, blotted to nitrocellulose and probed with $[^{32}P]$-nick-translated pA34. By establishing the sensitivity of this assay, using a range of prostate p(A^+) RNA concentrations, it was hoped to place an upper limit on the possible level of C3(2) mRNA. Indeed, the technique proved very sensitive, detecting C3 transcripts in 0.1 ng p(A^+) RNA, but it was not possible to resolve the 171 b and 147/149 b bands on the agarose gel system. Further difficulties were experienced due to the relatively low blotting efficiency of these small fragments.

Consequently any estimate of the maximum possible levels of C3(2) mRNA in rat ventral prostate must rely on the results from acrylamide gels. When S1 mapping the 5' ends of C3 transcripts, they were easily detected in 0.01 µg p(A^+) RNA (Fig. 13B, track 3) and were also detected in 0.001 µg p(A^+) RNA in the primer extension mapping experiment (Fig. 13A, track 4). However, the production of protected bands of 147/149 b using the C3(1) Xba I-Alu I probe was not observed even when using 10 µg p(A^+) RNA, suitable experimental conditions (Fig. 20B) and long gel exposures. This means that C3(2) transcript levels are probably below 0.01% (0.001 µg x 100/10 µg) of C3(1) mRNA levels.

R5: CHARACTERIZATION OF THE C3 ALU-LIKE REPETITIVE ELEMENT

5.1 Introduction

The Alu-like repetitive element found upstream of both C3 genes appears to be an example of the rat middle repetitive sequences which are found interspersed amongst single copy DNA. The resulting DNA organization, with repetitive sequences juxtaposed to protein coding regions, suggested to Davidson and Britten (1979) that repetitive DNA may be involved in the regulation of transcription from adjacent genes. Originally this theory was proposed to explain the changing patterns of expression of 'housekeeping' genes in various differentiated cells. Genes encoding proteins such as globin and indeed rat prostatic steroid binding protein, which are expressed to extremely high levels in only one cell type, were assumed to have their own specific regulatory mechanisms. Thus, the regulation of C3 expression by androgens would fulfil this criterion and an involvement in C3 expression by repetitive sequences (or their transcripts) would not be anticipated. However, the presence of this Alu-element so close to the start of C3 transcription is very intriguing in the light of Davidson and Britten's theories.

Consequently this section of results begins by describing the identification and characterization of the C3 Alu-like repetitive element and proceeds to examine the case for its possible functional involvement in C3 expression.

5.2 Identification of the C3 Repetitive Element

The repetitive element (RE), present between -450 and -370 within the 5' flanking DNA of the C3 genes, was initially identified by comparison with a library of published sequences using computer analysis. This revealed that the RE shows strong homology to Alu-like sequences isolated from rat tissue as two genomic and two cDNA clones, as illustrated in Fig. 21. These other examples of the RE originated as follows:-

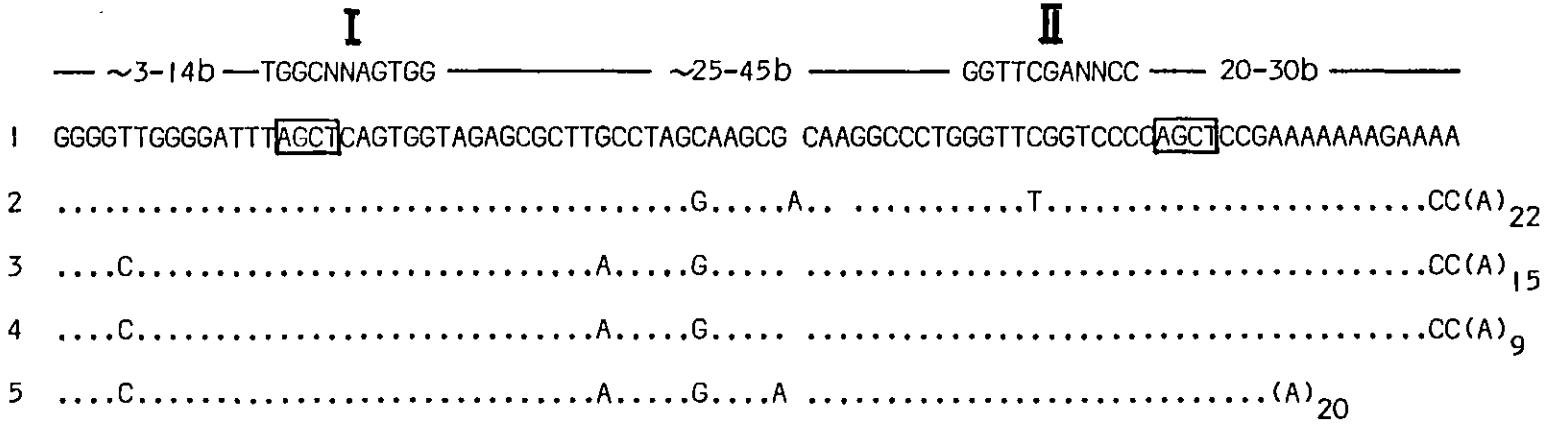
Sequence 2 was found interrupting a rat α-tubulin pseudogene which otherwise resembled an insertion of mature α-tubulin mRNA into the genome (Lemischka and Sharp, 1982).

Sequence 3 was found within the second intron of the rat growth hormone gene isolated by Barta et al. (1981). The intron in fact

Figure 21. A comparison of five short, Alu-like rat repetitive elements.

The Alu-like sequence found between -450 and -370 in both C3 genes is given in full in line 1. Alu I restriction sites are boxed and the space in the sequence has been introduced to help align four other examples represented below. A dot indicates these sequences are identical to line 1, while nucleotide differences are written in. Line 2 represents a sequence found within a rat α-tubulin pseudogene (Lemischka and Sharp, 1982). Line 3 represents a sequence found within the second intron of a rat growth hormone gene (Barta et al., 1981).
Lines 4 and 5 represent sequences isolated from rat brain p(A^+) RNA (Sutcliffe et al., 1982).

The top line in the figure shows the general dimensions of genes transcribed by RNA polymerase III (Galli et al., 1981).
Consensus sequences for the split 'promoter' or polymerase binding regions, termed box I and box II, are given in full above matching sequences within the five Alu repeats.


I
II
— ~3-14b — TGGCNNAGTGG ———— ~25-45b ———— GGTTCGANNCC —— 20-30b ———
1 GGGGTTGGGGATTTAGCTCAGTGGTAGAGCGCTTGCCTAGCAAGCG CAAGGCCCTGGGTTCGGTCCCCAGCTCCGAAAAAAAGAAAA
2 .....G.....A.. .....T.....CC(A)22
3 ....C.....A.....G..... .....CC(A)15
4 ....C.....A.....G..... .....CC(A)9
5 ....C.....A.....G....A .....(A)20

contained three repetitive elements: a tandem repeat of a rodent type-2 Alu sequences, followed by this shorter Alu-like element.

Sequences 4 and 5 were found at the 3' ends of two cDNA clones randomly isolated from rat brain p(A^+) RNA by Sutcliffe et al. (1982). The RE was the only common sequence between the two cDNA's and was also the only portion to hybridize to three other similarly isolated cDNA clones and to a 160 b RNA found in brain, but not liver or kidney tissue.

The two genomic isolates (sequences 2 and 3) were found to be flanked by direct repeats of cell DNA which were unique to the location of the particular RE. The C3 RE can also be assigned direct flanking DNA repeats, though their exact position is unclear, as the surrounding DNA sequences are anyway very repetitious. If repeats of 7 bp or more alone are considered, then two possible pairs have been indicated in Figs. 10 and 24.

All five sequenced examples of this RE are ~80 bp long, contain two restriction sites for Alu I and also possess good consensus sequences for the bipartite internal promoter, or binding site for RNA polymerase III. Furthermore, all five sequences are very similar, showing 96% homology to one another, which is high for an Alu-like family. Computer analysis revealed that these short Alu-like elements only possess 62% homology to the 5' end of a Chinese hamster 166 bp type 1 sequence and 61% homology to the 5' end of a rat 190 bp type 2 element. The C3 RE would therefore appear to belong to a distinct class of Alu elements.

The C3(1) 5' deletion mutant clone #5 (-292A) which contains the RE, plus DNA sequences further upstream (see Appendix), was radiolabelled by nick-translation and used to probe a series of Southern blots. This showed that the RE is not present within either of the genomic clones isolated for the other components of prostatic steroid binding protein, C1 and C2, and is not found anywhere else within the C3 genomic clones, including the 3' clone λ5D. When Eco RI cell DNA digests were probed a smear of hybridization down each track was observed, characteristic of interspersed repetitive sequences such as Alu families. Thus, the Alu-like RE within the C3 5' flanking DNA is probably represented in the haploid rat genome to the order of 10^4-10^5 copies (Jelinek and Schmid, 1982).

5.3 RE Expression

As the C3 RE possesses sequences that may be sufficient to render it an RNA polymerase III template, it is conceivable that the RE is expressed in vivo. Certainly, if Southern blots bearing this sequence are probed with either $[^{32}P]$-cDNA to prostate p(A^+) RNA or $[^{32}P]$-5' end labelled total prostate RNA, then hybridization to the RE is observed. Thus, expression of this, or similar repetitive DNA sequences must occur in ventral prostate tissue.

To try to characterize these transcripts an RE specific primer was prepared as described in Fig. 23. This 115 b primer, whose 3' end mapped 51 b short of the 5' end of the RE was used to assay p(A^+) RNA from rat liver, seminal vesicle and prostate tissue and also total prostate RNA (Fig. 22, tracks 2-6 and 12). The result was a complicated series of bands, presumably equivalent to the 5' ends of a number of RE-related transcripts, that appeared to be expressed in all the rat tissues examined.

In an attempt to study transcription from the C3 RE alone, the same primer was used to assay cytoplasmic RNA isolated from various lines of mouse S115 cells stably transformed with cloned C3(1) (50, C5, SV2 and SV6 cells) or C3(2) (A1 cells) DNA (see section M1, 1.8 for a full description). The S115 cells themselves and the C5 cell line were found not to contain RE transcripts (Fig. 22, tracks 7-9), but RNA from the other lines produced a distinct gel band corresponding to a primer extension of 45-50 b (Fig. 22, tracks 10 and 11; Fig. 23A, tracks 4 and 5). This implies that the C3 RE can be expressed in heterologous cells with the transcripts initiating at or near its 5' boundary (Fig. 24). However, the level of this expression, compared to C3 transcription within S115 cells, seemed disproportionately high and did not reflect the relative levels observed for these two transcripts in prostate tissue. For example, a primer extension assay of C3 and RE transcript levels in 50 cells, A1 cells and prostate is shown in Fig. 23A. From densiometric measurements and consideration of the amount of RNA used in each track it is apparent that transfected S115 cells produce 20 times more RE than C3 transcripts, while prostate RE expression is at most 1/600th of C3(T_1) mRNA levels.

Nevertheless, RE expression was further assayed by using S1 mapping to detect the 3' ends of the transcripts, using a 360 b probe (Fig. 23). Specific termination sites could not be mapped using ventral

Figure 22. Assaying for Alu transcripts in a variety of rat tissues and C3 transformed S115 cell lines.

All samples contained 30,000 c.p.m. 115 b primer described in Fig. 23 and 20 µg RNA comprising yeast carrier RNA plus: track 1, control; track 2, 5 µg seminal vesicle p(A^+) RNA; track 3, 5 µg liver p(A^+) RNA; track 4, 10 µg prostate p(A^+) RNA; track 5, 5 µg prostate p(A^+) RNA; track 6, 1 µg prostate p(A^+) RNA; track 7, 20 µg S115 cell RNA; track 8, 20 µg C5 cell RNA; track 9, 20 µg RNA from C5 cells cultured for three days without serum or testosterone; track 10, 20 µg 50 cell RNA; track 11, 20 µg A1 cell RNA; track 12, 5 µg total prostate RNA. Marker tracks M1 and M2 were 5' end-labelled Hpa II digested and Hinf I digested pAT 153 respectively.

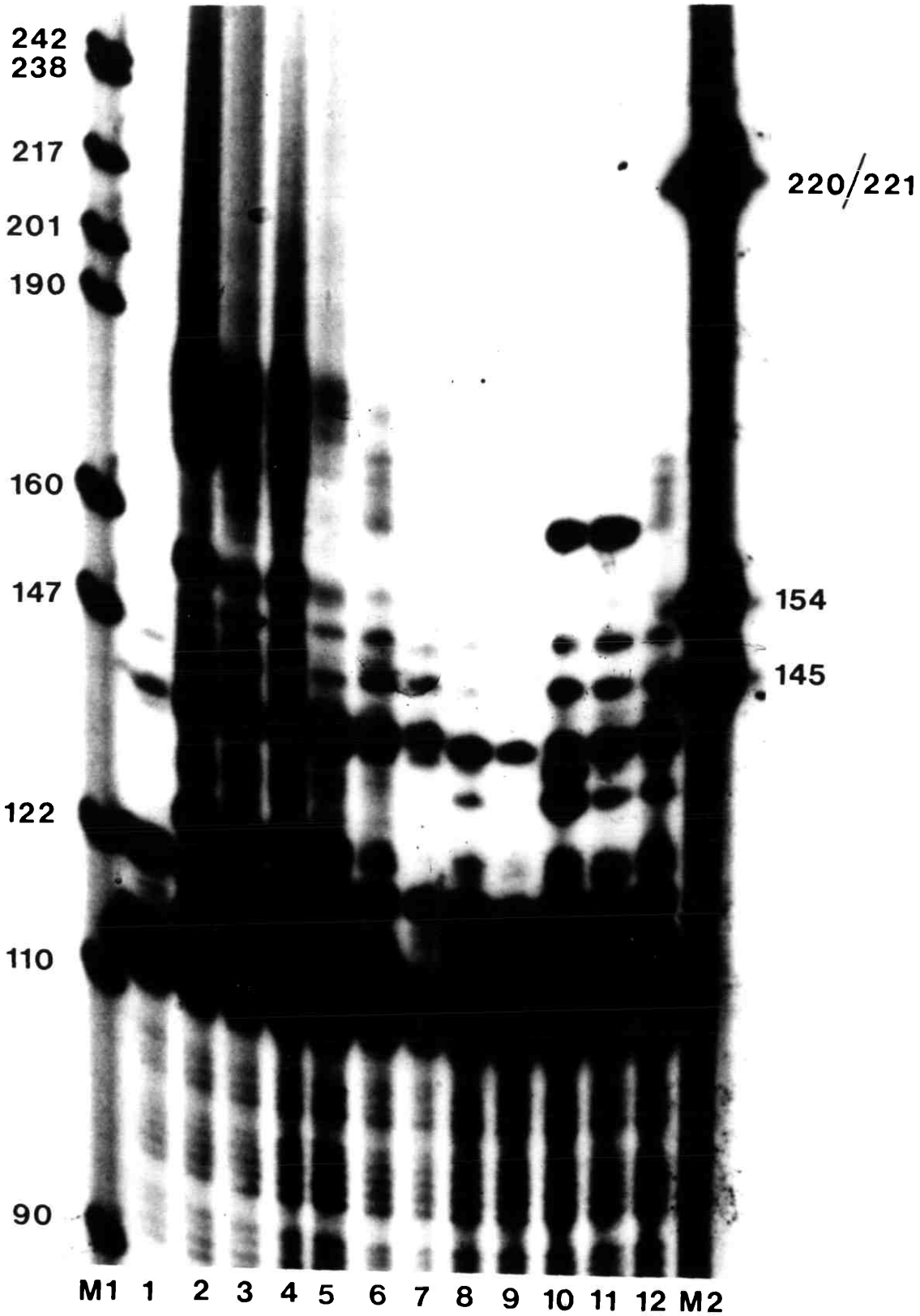

242
238
217
201
190
160
147
122
110
90
220/221
154
145
M1 1 2 3 4 5 6 7 8 9 10 11 12 M2

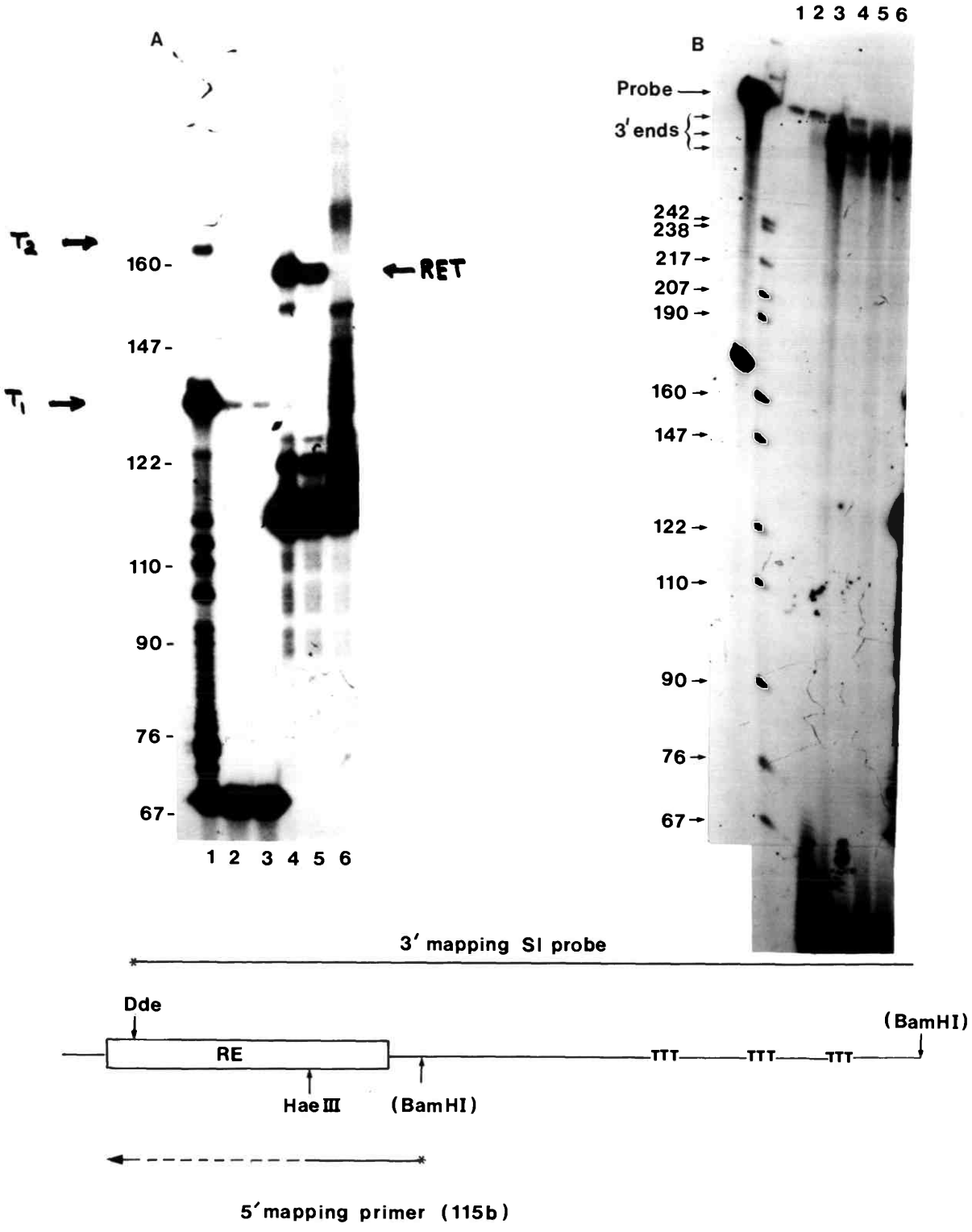
Figure 23. Mapping 5' and 3' termini of Alu transcripts.

Suitable probes and primers were prepared from the C3(1) 5' deletion mutants described in the Appendix. The 115 b primer for 5' end mapping was prepared from clone #5 (-292A) by 5' end-labelling at the Bam HI site, created at the deletion end point and redigesting with Hae III within the RE. A 3' mapping S1 probe was generated from clone #3 (-78A) by digesting with Dde I, resecting further with exonuclease III, filling in using $[\alpha^{32}P]$-dCTP and redigesting at the Bam HI site at the end of the deletion. This produced a 360 bp 3' end-labelled probe.

(A) Primer extension of C3 and RE transcripts from prostate and stably transfected cells was carried out using the 68 b C3 primer (tracks 1-3) or the 115 b RE primer (tracks 4-6). All samples contained 50,000 c.p.m. appropriate primer and 20 µg RNA comprising yeast carrier RNA plus: track 1, 0.5 µg prostate $p(A^+)$ RNA; tracks 2 and 4, 10 µg cytoplasmic RNA from 50 cells (C3(1)); tracks 3 and 5, 10 µg cytoplasmic RNA from A1 cells (C3(2)); track 6, 10 µg prostate $p(A^+)$ RNA. The 5' termini T_1, T_2 and RET (repetitive element transcript) are marked.

(B) Alu transcripts were assayed by S1 mapping of 3' ends using the 360 bp probe at 60,000 c.p.m. per track. RNA samples assayed were largely cytoplasmic RNA from stably transfected S115 cell lines. Track 1, control; track 2, 10 µg prostate $p(A^+)$ RNA; track 3, 30 µg RNA from A1 cells; track 4, 30 µg RNA from 50 cells; track 5, 30 µg RNA from SV2 cells; track 6, 30 µg RNA from SV6 cells.

Gel markers were 5' end-labelled Hpa II cut pAT 153.


A
T2
T1
160-
147-
122-
110-
90-
76-
67-
RET
1 2 3 4 5 6
B
1 2 3 4 5 6
Probe
3' ends
242
238
217
207
190
160
147
122
110
90
76
67
3' mapping SI probe
Dde
RE
TTT
TTT
TTT
(BamHI)
Hae III
(BamHI)
5' mapping primer (115b)

prostate p(A^+) RNA, but RNA samples from transfected S115 cells produced a series of discrete bands (Fig. 23B). The sizes of these protected fragments, determined from several experiments using different gel running times, were related to the DNA sequence. This showed that the RE transcripts terminated at clusters of three T residues within the single copy DNA extending beyond the Alu repeat towards the C3 cap sites (Fig. 24). Indeed, using a longer S1 probe some RE transcripts were found to terminate beyond T_1 and T_2. It should be noted that the intense band of 50-55 b in the gel tracks in Fig. 23B is not thought to represent a short family of RE transcripts, but is probably an S1 digestion artefact caused by duplex melting within the very A-rich region at the 3' end of the Alu element. The use of a lower digestion temperature (12^oC) alleviated, but did not wholly remove this problem.

As the RE is so close to the androgen-controlled C3 gene, it was conceivable that this sequence too might be expressed in an androgen-dependent manner. Certain of the stably transfected S115 cell lines show androgen controlled synthesis of C3 transcripts, as illustrated for SV2 and SV6 in Fig. 25B. From densiometric measurements of the gel autoradiograph these cell lines both exhibit a 3-4 fold stimulation of C3 transcript levels when cultured in the presence of 10^{-8} M testosterone. These same RNA samples were also assayed for RE transcripts by primer extension (Fig. 25A, tracks 1-4). In this case though the appropriate densiometric ratios were only 1.1 making it unlikely that androgens affect RE transcript levels.

Finally, some attempt was made to identify the RNA polymerase responsible for the observed RE transcription. The structure of the transcripts is reminiscent of polymerase III products, as might be expected with a promoter region for this enzyme apparently present within the RE. However, ~30 bp upstream of the start of RE transcription there is a sequence 5' TATTAA 3' (Fig. 24) which might function as a polymerase II promoter. To investigate these possibilities the RE was transcribed *in vitro* with appropriate concentrations of α-amanitin and the RNA products were assayed by primer extension (Fig. 25A, tracks 6-8). The RE seemed to be transcribed accurately *in vitro*, but low levels of α-amanitin caused some reduction in synthesis. This was coupled with the appearance of non-specific gel bands which persisted at high levels of α-amanitin and probably represent 'end to end' transcription of the RE containing restriction fragment. High levels

Figure 24. Features of the Alu element within the C3 genes.

The figure shows 5' flanking sequences of C3(1) around the Alu repeat. The putative promoter regions are boxed, either in full for parts I and II of the RNA polymerase III promoter, or dotted for the 'TATA' like sequence. Possible RE flanking direct repeats are shown as pairs of arrows, x and y. The site of transcription initiation is marked RET, while apparent transcription termination sites have been mapped to areas marked (┌──┐). The sequence of a short peptide that may be encoded by these transcripts is indicated. The nucleotide numbering corresponds to that given in Fig. 10 and represents distance from the C3 T_1 cap site.


RET
I
-450
-400
GAAAACTAGGGATGCTATTAACAAAGCAAATAAAGAGAGAAAAGAGGGGTTGGGGATTTAGCTCAGTGGTAGAGCGCTTGCCTAGCAAGCGCAAGGCCCT
x
y
II
-350
-300
GGGTTCGGTCCCCAGCTCCGAAAAAAAGAAAAGGAAAAAAAAAGAGAGAGAGAGAAAAAAGAATGAGGAGAATGGAGGAAAAGGAAGGGAGAGATGAGACC
x
y
METArgArgMetGluGluLysGluGlyArgAspGluThr
-250
-200
ACCTTAAAAATCAGAAGAGCAGATGATAAAACCAGCCCACAGTGCTGGAGTCTGCACTAAACTTTAGAAGTCAACACTTTGCTGTGGCCATTTGTCATCA
ThrLeuLysIleArgArgAlaAspAspLysThrSerProGlnCysTrpSerLeuHis...
150
CAGTGGTTCCCAGAGTCTCAAAATCTGTTTCCCTGCTTTCTTCAATGAAACCAGTGTTCTTTGGCTCTTCTTCGCCCTCTTATGTTT

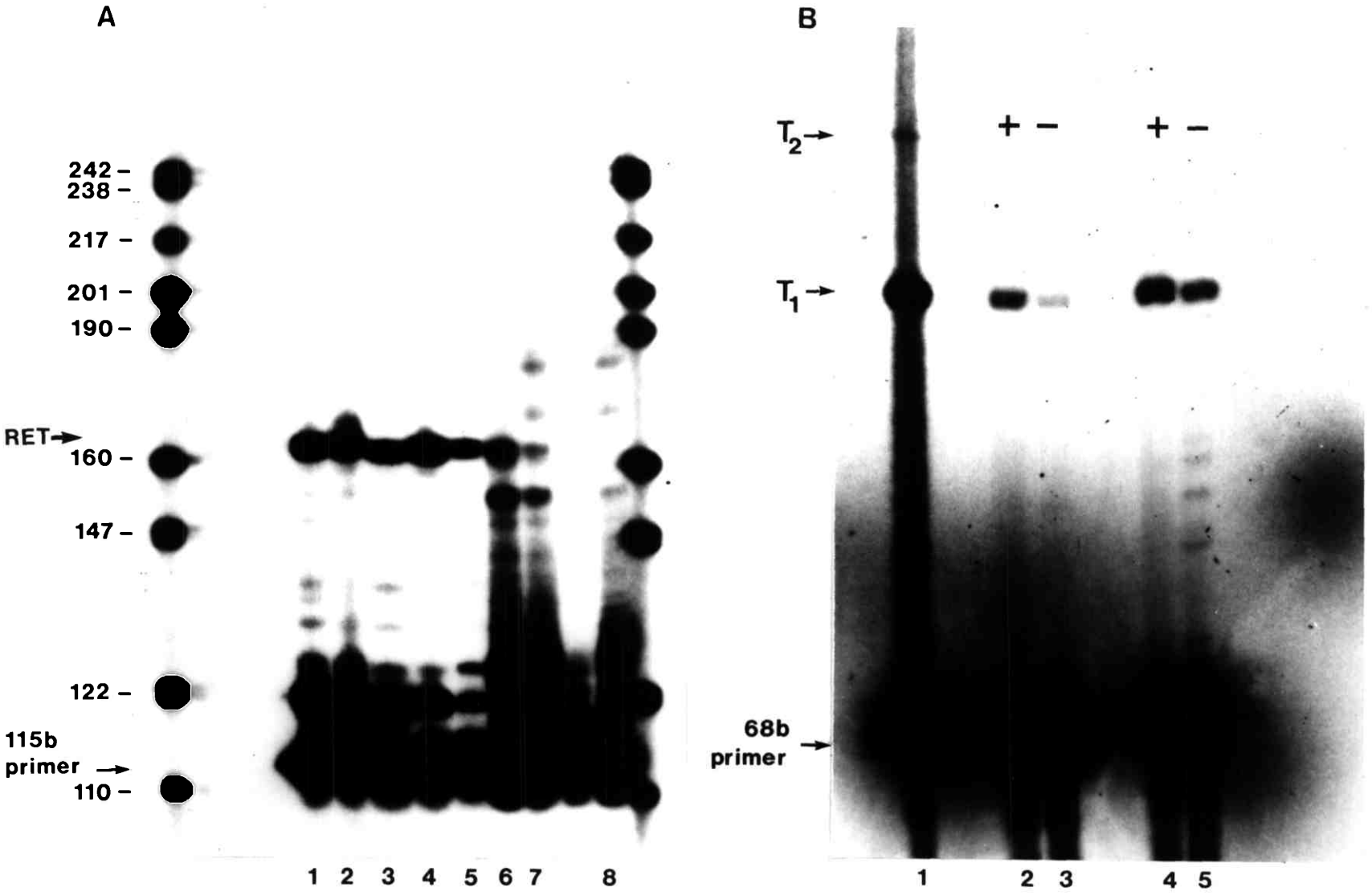
Figure 25. Further analyses of transcription from the Alu element.

(A) Two C3(1) stably transformed cell lines, SV2 and SV6, were maintained for 4 weeks in HEPES buffered E4 with 1% steroid depleted serum alone (-), or supplemented with 10^{-8}M testosterone (+). Cytoplasmic RNA (15 µg) from these cells was assayed by primer extension using the 115 b RE primer at 25,000 c.p.m. per track as shown in tracks 1(-) and 2(+) for SV2 and tracks 3(-) and 4(+) for SV6.

In vitro transcription mixes (100 µl) were supplemented with 40 µg/ml DNA from clone #4 (-116A) (see Appendix) which had been digested with Bam HI and Sfa NI (cuts 40 bp 5' of the Alu element). Primer extension analysis (using the 115 RE primer at 25,000 c.p.m. per track) of the RNA products from mixes containing 0 (track 6), 1 (track 7) and 200 (track 8) µg/ml α-amanitin is shown with track 5, a control containing 10 µg cytoplasmic RNA from 50 cells.

Gel bands corresponding to the 5' terminus of the Alu transcripts are marked RET.

(B) Primer extension analysis of levels of C3 transcripts within the hormonally manipulated SV2 and SV6 cells described above. All samples contained 30,000 c.p.m. of the 68 b C3 primer and RNA as follows: tracks 2-5, 15 µg cytoplasmic RNA from SV2 (tracks 2 and 3) or SV6 (tracks 4 and 5) cells; track 1, 1.0 µg prostate p(A^+) RNA plus 14 µg yeast carrier RNA.


A
242 –
238 –
217 –
201 –
190 –
RET
160 –
147 –
122 –
115b
primer
110 –
1 2 3 4 5 6 7 8
B
T2
T1
+ – + –
68b
primer
1 2 3 4 5

of α-amanitin did abolish RE transcription, however, indicating that this sequence can be trancribed by polymerase III, though some polymerase II transcription may also occur, particularly *in vitro*.

5.4 Discussion

The Alu-like C3 RE is clearly related to sequences expressed in a variety of rat tissues and can be transcribed from cloned DNA introduced into heterologous cells or into cell-free transcription assays. The transcriptional orientation is the same as for the adjacent C3 sequences, and the RE transcripts are characteristic of RNA polymerase III products. Initiation is at a unique site close to the first intragenic promoter sequence (box 1) and is roughly coincident with the beginning of the Alu element. Termination occurs within 3T clusters (usually in GC-rich regions) in single copy DNA beyond the repeat. A great variety of 3' ends have been found which may be accounted for by the observation that 3T clusters generally make weak polymerase III transcription terminators (Bogenhagen and Brown, 1981). Furthermore, these transcripts continue to be synthesized *in vitro* in the presence of low levels of α-amanitin, which would inhibit polymerase II, but not polymerase III (Weinmann *et al.*, 1974). As shown in Fig. 24, such transcripts could encode a small, 32 residue polypeptide (if the possibility of splicing is ignored), though this seems unlikely as the uncapped polymerase III transcripts are not thought to be recognized by the cellular translation apparatus.

Attempts to relate these observations on RE transcription to the ventral prostate, with a view to finding a functional role for this Alu element, were not very rewarding. A primer extension assay of *in vivo* RE-like transcripts revealed a confusing series of bands, but since the RE is highly reiterated in the genome this multiplicity was not really surprising. Clearly the degree and manner of expression from each Alu member will depend on its location within the genome. Some may be exclusively polymerase III templates, while others may be transcribed by polymerase II as part of otherwise single copy transcripts, like the RNA's which were cloned from rat brain (Sutcliffe *et al.*, 1982). Moreover, the primer used in the assay could conceivably cross-hybridize to sequences transcribed from other, related families of repeats. Thus it would seem impossible to make meaningful deductions about a possible functional role for the RE, particularly in respect of C3 transcription, but certain points should be noted:-

i) The pattern of bands found from primer extension analysis of rat liver, seminal vesicle and ventral prostate p(A^+) RNA was similar in all tracks (Fig. 22, tracks 2-6). In none of these tissues was there an appreciable population of RE transcripts of the type found *in vitro* or in transfected S115 cells (Fig. 23A and B).

ii) Clearly transcription from the RE is not susceptible to regulation by androgens (Fig. 25A, tracks 1-4).

iii) Overproduction of RE transcripts compared to C3 transcript levels only occurred in those cell lines known to possess multiple copies of the transfected DNA, e.g. A1 cells (Page and Parker, 1983). Cell lines with few copies of cloned C3 DNA (e.g. C5 cells) did not produce detectable levels of RE transcripts. It is therefore likely that within the stably transformed cell lines the C3 and RE transcripts were being produced from different individual copies of transfected cloned DNA.

If these Alu transcripts had a functional role in C3 expression *in vivo*, for example assisting the passage of transcripts to the cytoplasm, then one might expect roughly equivalent levels of C3 and RE transcripts in the ventral prostate. However, taken together, the points listed above show that this is not observed. Indeed, RE and C3 transcripts are clearly not co-ordinately regulated and are probably never jointly copied from the same DNA template, making it unlikely that the particular RE upstream of the C3 genes is expressed *in vivo*.

SECTION D: DISCUSSION

A sound knowledge of the structure of a gene is a necessary prerequisite to an understanding of the functional mechanisms which control its expression. This thesis has concentrated on characterizing the genes for the C3 component of rat prostatic steroid binding protein. It is therefore interesting at this stage to examine the results from this structural analysis in terms of the clues obtained towards understanding more functional problems associated with the androgen controlled expression of this polypeptide. To this end the results just presented are summarized and then discussed in wider terms than was appropriate at the end of each results section. The most interesting functional aspects are highlighted for the overall gene structure, the control of expression for both C3(1) and C3(2) and the possible role of the Alu-equivalent repetitive element. Where possible, ideas for extending particular areas of the project by further experimentation are also included.

1. Structure of the C3 Genes.

Using rat DNA λ-libraries it has been possible to isolate two non-allelic genes, C3(1) and C3(2), for the C3 polypeptide of prostatic steroid binding protein. These two genes have been found to be very similar in overall length (3.04 $\pm$ 0.2 kbp), restriction enzyme map and exon/intron structure. Their sequence determination has further revealed that, although overall they share >95% homology, only one of these genes, C3(1), is totally compatible with a C3 cDNA clone sequence (Parker et al., 1983) and with the C3 polypeptide sequence (Peeters et al., 1981).

Both genes possess two introns (of 1.7 kbp and 0.69 kbp) which seem to divide the coding regions so as to give three short exons encoding distinct functional units. The first exon encodes the complete signal peptide to ensure secretion of the C3 polypeptide. The major polypeptide sequences are encoded in the second exon, while the third exon contains sequences expressed as non-translated mRNA. The nucleotide differences between the expressed regions of the two genes are confined to 8 bp within the second exon, and 4 bp at the end of the third exon.

The high degree of homology between the two genes extends into the flanking DNA and even to the intervening sequences. This suggests that the two genes resulted from the duplication of an ancestral C3 gene. However, it would be unwise to speculate as to the evolutionary age of this duplication, as has been estimated for other multigene families such as the β-globin genes (Efstradiadis et al., 1981). Conceivably, the C3 duplication occurred fairly recently in evolutionary time, but the extent of the sequence homology between all regions of the genes implies that a process of gene conversion (reviewed Baltimore, 1981) may have occurred, thus correcting one gene sequence against the other some time after the ancestral duplication.

In this context it is interesting to note that the same overall gene structure is found for the genes to the other components of prostatic steroid binding protein, C1 and C2. These two genes are similar in length to the C3 genes (~3.2 kbp) and also have three short exons separated by two introns of 1.7 to 1.8 kbp and 0.8 to 0.9 kbp respectively. Again, also the first exon encodes the signal peptide (Parker et al., 1982b). Furthermore, sequence analysis of these two genes has revealed that they share >76% homology in their 5' flanking and coding regions, resulting in 50-60% homology at the amino acid sequence level (Parker et al., 1982a, 1982b). Thus, these two genes seem to have also arisen by gene duplication. Conceivably then, the ancestral C3 and C1/C2 genes may also have originated from an even older duplication. For this reason the sequences of all four present day genes were compared to look for homologies. No extensive regions of homology were found, but short regions of >75% similarity were observed between the C3 genes on the one hand and the C1 and C2 genes on the other. Therefore, nothing really definitive may be said about the evolutionary origin of these four genes. However, if they did all arise from the duplication of a single ancestor, then the present day genes might be expected to be found closely linked on one rat chromosome. Such a possibility may be examined by chromosome 'walking', though the genes must be at least 10 kbp apart as no overlapping clones were isolated from the libraries screened so far. For example, the clones λ11B and λ5D together cover ~22 kbp of rat DNA within which only C3(2) sequences can be recognized.

2. C3 expression in vivo.

Due to its compatibility with the C3 cDNA clone pA34 and the C3 polypeptide, the C3(1) gene must be the most active of the two genes in vivo. Transcript mapping studies have shown this gene to possess two transcription initiation regions, T_1 (major) and T_2 (minor), each of which is preceded by a 'TATA' box homology. The transcripts probably have a unique mature 3' terminus 12 bp downstream of a putative polyadenylation signal, AUUAAA. Thus, the population of in vivo C3 mRNA is made up of two families of transcripts with a common 3' end and 5' ends which map 36 bp apart and which occur at a constant ratio of 18:1, even within hormonally manipulated rats. The ratio of incidence of the two transcripts probably reflects the ratio of the rate of transcriptional initiation occurring at T_1 and T_2, although this has not been verified. Analysis of transcripts labelled in vitro from nuclei isolated from rat ventral prostate tissue may confirm this, though it may be difficult to separately evaluate transcripts from T_1 and T_2 with only 36 bp of differing sequence.

Using cloned DNA, C3(2) has also been found capable of producing transcripts from T_1 and T_2. Page and Parker (1983) showed that transfection of both C3(1) and C3(2) into the androgen receptor containing S115 cell line produced authentic C3 transcripts, responsive to androgenic control. As C3(1) and C3(2) possess only 1 base difference from -235 to within the first exon, this perhaps was not surprising. Attempts presented in this thesis to directly compare the transcriptional efficiency of cloned DNA from both genes in an in vitro assay were largely frustrated by the poor specificity shown by both templates in the assay (section R4). Authentic initiations at T_1 and T_2 were largely hidden amongst a whole series of heterologous ones. A comparison of the transcriptional efficiencies of C3(1) and C3(2) might be better achieved by their separate introduction into heterologous cells in a short-term transfection assay. Co-transfection of a distinct gene (for example the Herpes Simplex virus thymidine kinase gene) could act as an internal control against which the relative amounts of expression from the two C3 genes may be measured.

Nevertheless, as C3(2) can clearly produce authentic C3 transcripts in transfected heterologous cells then it is possible that it is expressed to some degree within the rat ventral prostate. However, from several lines of evidence it would seem that the C3(2) gene is in fact transcribed poorly, if at all in vivo.

i) Protein sequence. There is no evidence from the polypeptide sequencing data (Peeters et al., 1981; Dr. W. Rombauts, personal communication) that a C3(2) polypeptide exists. Any such polypeptide forming 5% or more of total C3 secreted from the ventral prostate would have been detected, although it might be argued that C3(2) could be transcribed, but not translated.

ii) R-looping. Although R-loops were identified between prostate $p(A^+)$ RNA and the C3(1) clone, λ6, similar structures using the C3(2) clone, λ5A, were never detected (Fig. 7). Instead, only incomplete R-loops were observed in which the second exon of C3(2) failed to hybridize, presumably due to the sequence mismatching in this region between C3(2) and C3(1) mRNA. Examination of 30-40 such incomplete R-loops never revealed a case of second exon hybridization as would be expected were homologous C3(2) transcripts present in ventral prostate. Though inconclusive, this observation again suggests that C3(2) mRNA represents less then 5% of C3 mRNA.

iii) Sequence of transcripts from T_2. It was noted in Section R3 that primer extension sequencing of C3 transcripts originating at T_2 revealed a 5' end sequence that could only have been encoded by C3(1). Any C3(2) transcripts present at levels detectable in the experiment in Fig. 14B (probably 1-2% of C3(1) mRNA levels) would have given a faint band in the C track level with the middle band of the T triplet. The absence of such a band indicates that it is unlikely that transcripts from C3(2) T_2 form a significant population of ventral prostate C3 mRNA. However, this of course gives no indication of the possible levels of transcription from C3(2) T_1.

iv) Direct assays for C3(2) mRNA. Due to the great similarity between the two C3 gene sequences it was extremely difficult to directly assay for in vivo C3(2) transcripts alone. Inspite of this the results presented in Section R4 indicate that C3(2) is probably expressed at levels below 0.01% of C3(1) and therefore can only be present at about 0.001% of total $p(A^+)$ RNA at most.

It would therefore appear that the C3(2) gene is virtually silent in vivo, though it might be argued that the gene is in fact expressed at normal rates within the ventral prostate, but that the transcripts are extremely labile. Although this cannot be ruled out, it is rather unlikely because the C3(2) transcripts were clearly stable within the 'foreign' environment of the S115 cells.

An indication that C3(2) is indeed silent *in vivo* was provided by White and Parker (1983). They examined the state of DNA methylation at CCGG sites around all four prostatic steroid binding protein genes using the pair of isoschizomeric restriction enzymes Msp I and Hpa II. Methylation at the internal C residue inhibits Hpa II digestion. A summary of the results they obtained for the two C3 genes is presented in Fig. 26. It can be seen that, at the sites assayed, the C3(2) gene exhibits a constant high degree of methylation. C3(1) is also highly methylated within most rat tissues, but during ventral prostate development this gene becomes demethylated, so in the mature animal C3(1) appears hypomethylated compared to C3(2).

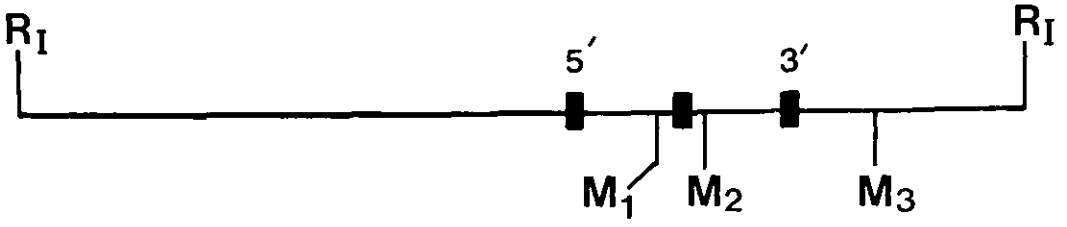
These observations are extremely interesting in the light of theories proposed by Razin and Riggs (1980) that changes in DNA methylation at CpG sites may control gene expression in a heritable manner in most eucaryotic systems. In general, the transcriptional activity of a vertebrate gene or virus can be inversely correlated with its state of methylation (reviewed Felsenfeld and McGhee, 1982; Doerfler, 1983). That this is not always so has lead to criticism of the theory as changes in gene expression are sometimes observed to precede demethylation (Kunnath and Locker, 1983). Consequently methylation changes may be the result of, rather than the cause of changes in gene transcription rates. However, by using restriction enzymes to analyse the degree of methylation of a gene, it is sheer chance if thereby the state of methylation of sequences important to the control of gene expression is also assayed. Recent studies by Busslinger *et al.* (1983b) may prove more definitive. They were able to methylate *in vitro* defined regions of the human γ-globin gene and examine the effect on expression in stably transformed mouse L-cells. This showed that methylation of globin coding regions alone had no adverse effect on gene transcription. However, if the 5' flanking DNA (from -760 to +100), including putative promoter regions, was methylated then γ-globin gene expression became undetectable.

Whether or not the hypermethylation of C3(2), compared to C3(1), is the cause or effect of its lack of *in vivo* expression it probably does indicate that this gene is truly silent within the ventral prostate. This would fit in well with the finding that cloned C3(2) DNA, that is therefore unmethylated at CpG sites, is transcriptionally active in heterologous cells (Page and Parker, 1983) and *in vitro* (Section R4). Conceivably then the methylated sites assayed by Msp I/Hpa II digestions

Figure 26. Summary of the state of DNA methylation within and flanking the two C3 genes.

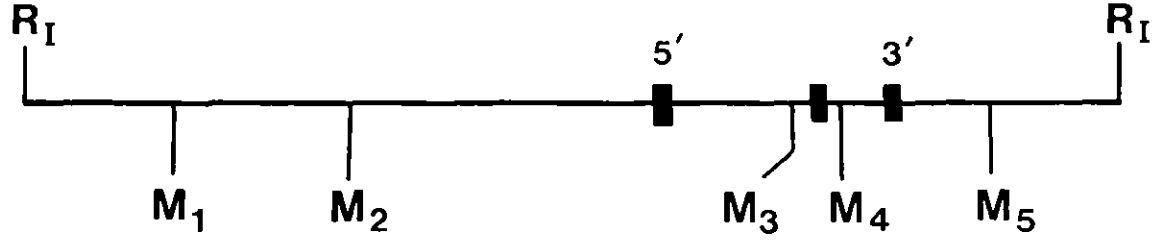
Simple restriction maps of C3(1) and C3(2) are given, showing Eco RI (RI) and Msp I (M) sites only, with exons blocked in. The state of methylation at the numbered Msp I sites was examined in a variety of rat tissues (seminal vesicle, coagulating gland, heart, liver, kidney, spleen, sperm and ventral prostate from normal and androgen-withdrawn rats). The degree of methylation at each site was estimated from Southern blots and is shown as: 0 = 0%, + = 50% and ++ = 100% methylated. The asterisk (*) for M2 of C3(2) indicates that this site was not digested by either Hpa II or Msp I which usually means the enzyme recognition site ($C^{M}C$ G G) is within the inhibitory hexanucleotide sequence G G $C^{M}C$ G G (Busslinger et al., 1983a). The diagram and results were taken from White and Parker (1983).

C3 (1) (λ6)



Ventral prostate from 10 day old rats	++	++	++
Ventral prostate from normal rats	0	0	0
Ventral prostate from withdrawn rats	+	+	+
Other tissues and sperm	++	++	++

C3(2) (λ11B)



All rat tissues and sperm	+	++*	++	++	++

are crucial to the expression of this gene *in vivo*. This could perhaps be examined by modifying cloned DNA (from C3(1) and C3(2)) either by using the bacterial enzyme Hpa II DNA methyl-transferase (Vardimon *et al.*, 1982) or using cloning vectors that can replicate both in *E.coli* and *H.parainfluenzae* (Danner and Pifer, 1982). Methylated and unmethylated cloned DNA may then be compared for transcriptional activity perhaps by microinjection into the nuclei of Xenopus oocytes. This system has been used extensively for such assays mainly because *in vitro* systems do not differentiate between methylated and unmethylated DNA (reviewed Doerfler, 1983).

If methylation is important for the control of C3 gene expression then there must be some mechanism whereby C3(1) alone is programmed for demethylation during prostate development while the apparently similar C3(2) gene is 'ignored'. Conceivably this could be a specific signal sequence present only in C3(1), but the degree of homology between the two genes makes it unlikely that this resides very close to the protein coding regions. Certainly, with only 1 bp difference between nucleotide -235 to within the first intron, such a regulatory sequence cannot be within the proximal 5' flanking DNA. Another possibility is that C3(2) remains silent because it is in an adverse chromosomal position. For example, the C3(2) gene may always be condensed within heterochromatin and therefore remain inaccessible to the demethylation enzymes and the transcriptional apparatus. Some very preliminary blotting data showing that the DNA flanking C3(2) is particularly highly reiterated in the rat genome is perhaps consistent with this notion.

3. The Alu-like repetitive element.

The only repetitive DNA sequence around the two C3 genes studied in any detail was the small Alu-like element found between -450 and -370 in both genes. Clearly this element and the other published similar sequences belong to an Alu family as their structural organization follows the same rules listed in the Introduction. There is a clearly defined 5' end for which a consensus sequence may be written, an A-rich 3' end and all the genomic isolates are flanked by unique direct repeats of cell DNA of at least 7 bp. The sequence homology for these Alu elements is very high (96%). However, the similarity at a sequence level to rodent type 1 and type 2 Alu-equivalent families is poor, as the consensus sequences of both of these families are at least

twice as long and show only ~60% nucleotide homology to these shorter elements. For these reasons it is suggested that the Alu element upstream of the C3 genes is a member of a distinct new rodent type 3 Alu-equivalent family.

The high sequence homology between the genomic isolates of the type 3 family, which are from apparently disparate locations, suggests either that there is some evolutionary constraint on sequence divergence, or that this family has evolved recently and rapidly dispersed throughout the genome. An evolutionary constraint would further imply that these elements perform some functional role and this possibility has been examined for the individual case of the C3 genes.

This family of Alu elements carry internal sequences compatible with the bipartite intragenic promoter or binding site for RNA polymerase III. Thus, these elements may be expressed in vivo and perform functions, suggested for other Alu-like transcripts, at the level of processing and transport of specific protein coding mRNA's. For such a role to be suspected though, there should be an observable correlation between expression from the Alu element and the adjacent C3 gene. However, as described in Section R5, the C3 Alu element does not appear to be transcribed in vivo although it is clearly capable of producing transcripts from cloned DNA templates in vitro and in heterologous cells. Even within whole cells though the C3 gene and its adjacent Alu sequence are seemingly not co-ordinately expressed, as discussed in Section R5. Hence it is doubtful whether these Alu transcripts could have a functional role in C3 expression. Alternatively, the DNA sequence itself may influence local chromosome structure and therefore transcription from the C3 gene, although this too seems unlikely. C3 expression is assumed to be co-ordinately regulated with the genes encoding the other components of prostatic steroid binding protein, namely C1 and C2, as all three gene products must come together to form a single final protein. Certainly, transcription from all three genes was influenced by androgens to a similar degree in isolated prostate nuclei (Page and Parker, 1982). However, blotting experiments revealed that neither of the C1 or C2 genes contain examples of this Alu element in their flanking DNA or introns. Hence, as it is unlikely that C3 alone would be controlled by such a sequence, it seems that in the case of the C3 genes the adjacent Alu element plays no regulatory role in gene expression.

The alternative explanation for the existence of the Alu RE is that it has recently dispersed throughout the rat genome. In evaluating this idea several pieces of evidence showing apparent sequence mobility should be considered:

i) As described in R5, all the genomic isolates of this repeat family are flanked by direct repeats of cell DNA. Furthermore, the transcripts, that the C3 RE was found capable of producing in some circumstances, are characteristic of Alu transcripts as described in the Introduction. They could therefore return to the genome perhaps in the manner proposed by Jagadeeswaran et al. (1981) and illustrated in Fig. 1.

ii) The example of a type 3 element found within a rat α-tubulin pseudogene (Lemischka and Sharp, 1982) is intriguing. The tubulin pseudogene itself resembles a mature tubulin mRNA inserted into the cell DNA. It therefore belongs to a class of protein coding pseudogenes (processed pseudogenes) believed to represent the return of RNA to the genome (reviewed Sharp, 1983), being flanked by direct repeats of cell DNA itself. As RNA polymerase II transcription is controlled by sequences outside those copied into RNA then the transcripts that return to the genome must in general be silent and may be termed ignorant DNA, an expression first used by Dover (1980). The insertion of the Alu element within the tubulin pseudogene caused a duplication of 15 bp of this sequence as the direct repeat flanking the RE. This structure therefore appears to represent the insertion of a piece of selfish DNA (that may be expressed) into an ignorant DNA sequence.

iii) The type 3 Alu sequence within the rat growth hormone gene is located within the second intron following two copies of a rat type 2 sequence in tandem (Barta et al., 1981). This isolate of rat growth hormone gene was recovered from Sprague-Dawley rats, but the gene has also been isolated from Hooded rats (Page et al., 1981). In this second example of the gene the type 3 Alu sequence is precisely missing from the second intron, although the two type 2 elements are still present. It is therefore clear that the distribution of Alu type 3 sequences differs between different rat strains. Furthermore, as there appears to be only one growth hormone gene in the rat genome (Page et al., 1981), presumably the RE plays no functional role in its expression if it is not present within the gene in certain strains of rat.

Consequently it may be possible to definitively show that C3 gene expression is not influenced by the Alu sequence by isolating an active C3 gene from a strain of rat which does not carry the RE upstream of the C3 gene. Some blotting studies have been performed attempting to detect restriction fragment size differences that might indicate the absence of the Alu element in the 5' flanking DNA of the C3 genes. DNA from the Fischer and Wistar rat strains were examined in this regard and were both found to carry the type 3 sequence within their genomes, but it was not possible to say if a copy was located within the C3 genes. Improved DNA blotting might provide an answer to this question.

On the evidence presently available though it would seem that the type 3 Alu sequences behave as parasitic elements capable of independently replicating in the genome and thereby surviving the pressure of natural selection. This means that they can persist without fulfilling any useful function and represent so-called selfish DNA (Orgel and Crick, 1980; Doolittle and Sapienza, 1980).

As the Alu element is present in C3(1) and C3(2) it implies that it integrated into the proposed ancestral C3 gene prior to the duplication generating the present day genes. Alternatively, one C3 gene may have acquired the original insertion and both now bear it owing to the gene conversion process that may have occurred between the two genes. Gene conversion might also account for the high sequence homology observed between individual type 3 family members. The correction of one sequence against another might occur between different genomic copies, or alternatively between the genomic sequences and DNA copies of their transcripts as suggested to explain the apparent sequence correction of the 5' third of the two types of rat U3 sn RNA gene (Bernstein et al., 1983).

4. Control of C3(1) gene expression.

If the Alu element is a fortuitous intrusion into C3 5' flanking DNA, its presence has clearly not disrupted the androgenic control of C3 expression. This implies that if DNA sequences further upstream are involved in that regulation then they can act over a distance and are unaffected by the insertion of intervening DNA. More plausibly, any regulatory sequences within the C3 upstream DNA probably lie between the Alu element and the C3 cap sites, that is within the first 350 bp and 5' flanking DNA. This would agree well with findings from other steroid controlled genes where sequences upto -200 seem sufficient to

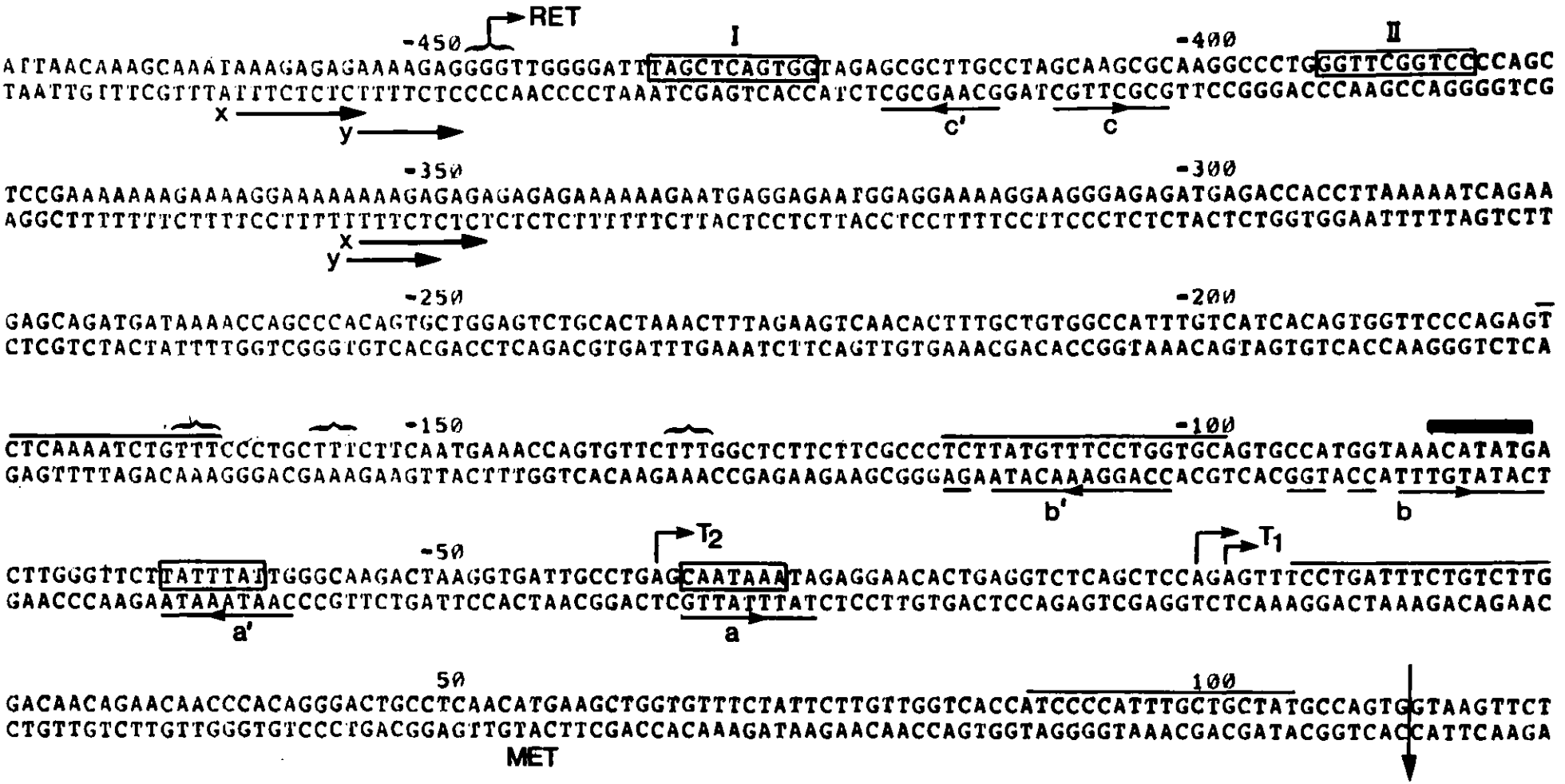
ensure hormonal regulation as discussed in the Introduction (see Table 2). With this in mind, C3(1) upstream sequences were examined for structures that might have a functional significance in androgen inducible C3 gene expression as discussed below and summarized in Fig. 27.

Both of the C3 transcription initiation regions, T_1 and T_2, are preceded by a 'TATA' box homology, but there is no sequence which shares extensive homology with the so-called CAAT-box that is found 70-90 nucleotides infront of several eucaryotic genes (Benoist et al., 1980). Other elements within gene flanking DNA which have been shown to be functionally significant in transcriptional regulation, are pairs of inverted repeats. Increasing numbers of genes transcribed by RNA polymerase II have been found to possess such repeats within the first ~100 bp of flanking DNA (reviewed Mirault et al., 1982). As shown in Fig. 27, the C3(1) gene seems to be no exception with the proximal 5' flanking DNA containing three pairs of inverted repeats located as follows: a) between -35 to -27 and -69 to -61 and therefore comprising both 'TATA' box homologies; b) an imperfect inverted repeat at -118 to -104 (TCtTATGTTTCCTGG) and -80 to -96 (CCAtGGtAAACATAtGA); and c) a closely hyphenated inverted repeat within the Alu element and centred at -413. It should be noted that repeats 'b' and 'c' also occur in C3(2), but repeat 'a' is not present due to the A ⟶ G transversion within the T_1 'TATA' box of C3(2).

The 685 bp of C3(1) 5' flanking DNA sequence available were also examined for long runs of alternating purine and pyrimidine residues. Such sequences favour the formation of a left-handed form of the double helix within supercoiled DNA which has been termed Z-DNA (Wang et al., 1979). Monoclonal antibodies to Z-DNA have recently been used to show that the SV40 72 bp repeat enhancer sequence contains three 'blocks' of Z-DNA, at least one of which was required for viral transcription and hence survival (Nordheim and Rich, 1983). Each of these DNA regions required only 8 alternating purine and pyrimidine resudies to form a stable Z-DNA structure. It was therefore interesting to note that the only region within the C3(1) 5' flanking DNA that might adopt a Z-DNA form was the sequence 5' (-87) ACATATG (-81) 3' which also coincides with one part of the long, imperfect inverted repeat (see Fig. 27). This 7 bp sequence might therefore influence C3(1) expression, particularly if the ability to form Z-DNA is stabilized by the binding of proteins, such as the S-Rc complex.

Figure 27. Features within the proximal 5' flanking DNA of the C3(1) gene.

The numbered sequence corresponds to that given in Fig. 10. The Alu element is indicated flanked by two possible pairs of direct repeats (x and y) and the bipartite RNA polymerase III promoter-like sequences within it are boxed (I and II). Possible sites of transcriptional initiation (RET) and termination (major ones only, ⌐⌒¬) are also given. C3(1) transcription initiation regions, T_1 and T_2, are marked and their corresponding TATA boxes boxed. MET indicates the start of C3 protein coding sequences and the end of the first exon is shown by a vertical arrow (↓). Other features indicated include 3 pairs of inverted repeats (a/a', b/b' and c/c'), a 7 bp region that may form Z DNA (▬) and four regions that show a high degree of homology to the proposed progesterone-receptor DNA binding site (Mulvihill et al., 1982), which are overlined.


RET
-450
I
-400
II
ATTAACAAAGCAAATAAAGAGAGAAAAGAGGGGTTGGGGATTTAGCTCAGTGGTAGAGCGCTTGCCTAGCAAGCGCAAGGCCCTGGGTTCGGTCCCCAGC
TAATTGTTTCGTTTATTTCTCTCTTTTCTCCCCAACCCCTAAATCGAGTCACCATCTCGCGAACGGATCGTTCGCGTTCCGGGACCCAAGCCAGGGGTCG
x
y
c'
c
-350
-300
TCCGAAAAAAAGAAAAGGAAAAAAAAGAGAGAGAGAGAAAAAAGAATGAGGAGAATGGAGGAAAAGGAAGGGAGAGATGAGACCACCTTAAAAATCAGAA
AGGCTTTTTTTCTTTTCCTTTTTTTTCTCTCTCTCTCTTTTTTCTTACTCCTCTTACCTCCTTTTCCTTCCCTCTCTACTCTGGTGGAATTTTTAGTCTT
x
y
-250
-200
GAGCAGATGATAAAACCAGCCCACAGTGCTGGAGTCTGCACTAAACTTTAGAAGTCAACACTTTGCTGTGGCCATTTGTCATCACAGTGGTTCCCAGAGT
CTCGTCTACTATTTTGGTCGGGTGTCACGACCTCAGACGTGATTTGAAATCTTCAGTTGTGAAACGACACCGGTAAACAGTAGTGTCACCAAGGGTCTCA
-150
-100
CTCAAAATCTGTTTCCCTGCTTTCTTCAATGAAACCAGTGTTCTTTGGCTCTTCTTCGCCCTCTTATGTTTCCTGGTGCAGTGCCATGGTAAACATATGA
GAGTTTTAGACAAAGGGACGAAAGAAGTTACTTTGGTCACAAGAAACCGAGAAGAAGCGGGAGAATACAAAGGACCACGTCACGGTACCATTTGTATACT
b'
b
T2
T1
-50
CTTGGGTTCTTATTTATTGGGCAAGACTAAGGTGATTGCCTGAGCAATAAATAGAGGAACACTGAGGTCTCAGCTCCAGAGTTTCCTGATTTCTGTCTTG
GAACCCAAGAATAAATAACCCGTTCTGATTCCACTAACGGACTCGTTATTTATCTCCTTGTGACTCCAGAGTCGAGGTCTCAAAGGACTAAAGACAGAAC
a'
a
50
100
GACAACAGAACAACCCACAGGGACTGCCTCAACATGAAGCTGGTGTTTCTATTCTTGTTGGTCACCATCCCCATTTGCTGCTATGCCAGTGGTAAGTTCT
CTGTTGTCTTGTTGGGTGTCCCTGACGGAGTTGTACTTCGACCACAAAGATAAGAACAACCAGTGGTAGGGGTAAACGACGATACGGTCACCATTCAAGA
MET

In the Introduction it was mentioned that some groups have identified sequences infront of genes which they believe to be important in the steroid inducible expression of those genes. Many of these sequences are very short and their importance is often deduced from their reiteration within the proximal 5' flanking DNA of the gene in question. In contrast, the 19 b consensus sequence postulated as being a chicken progesterone-receptor-DNA binding site (Mulvihill et al., 1982) was of sufficient length to allow the use of the computer to search for similar sequences within the C3 gene. Surprisingly, four regions showing significant homology to this consensus sequence were found around the cap sites of C3(1) (and C3(2)). These homologies ranged from 68% - 80%, values which rose to 74% - 83% when only the central 17 b of the consensus sequence were considered. These values are as good as many of the homologies found by Mulvihill et al. (1982) within the chicken egg white protein genes when compiling the consensus sequence. Furthermore, they show the same reiteration around the cap site. The exact locations of these homologies within the C3 genes are shown by overlining in Fig. 27 where it can be seen that one of the homologies overlaps with the other part of the inverted repeat 'b'. The degree of homology of each C3 region to the consensus sequence is shown in Table 3 which also illustrates that the C1 and C2 genes both contain two similar regions of homology around their cap sites.

As there is no evidence that progesterone modulates the expression of prostatic steroid binding protein, finding these homologous regions was highly unexpected. It would imply not only that S-Rc binding regions within regulated genes show some affinity irrespective of the particular steroid, but also that there is strong evolutionary conservation of these binding sites (from chicken to rat in this case). On the face of it this would seem unlikely as it does not equate with the observed specificity of action of different steroids. Intriguingly though, it has been noted that a sequence thought to be important in the poly (rI) - poly (rC) induction of the human interferon β gene is also remarkably similar ($^{13}/_{19}$ bp) to the progesterone-receptor binding site consensus (Tavernier et al., 1983). Conceivably then a certain class of DNA binding proteins involved in the induction of gene expression may recognize a core sequence 'motif' while other flanking sequences may modulate the binding by different proteins to generate the observed specificity. Indeed, it has been shown that the

Table 3. Regions within the 5' DNA of prostatic steroid binding protein genes showing significant homology to the consensus sequence for the chicken progesterone receptor-DNA binding site (Mulvihill et al., 1982).

Gene	Position	Consensus sequence A T C C/T C/T A T T A/T T C T G G/T T T G T A	x/19	%*	y/17†	%
C3(1) and C3(2)	-182 to -163	g T C T C A a a A T C T G T T T c c c	13	68	13	76
	-119 to -100	c T C T T A TgT T T CcT G G T G c A	16	80	15	83
	+4 to +23	t T C C TgA T T T C T G TcT T G g A	16	80	15	83
	+87 to +107	A T C C CcA T T T g C T GcT a T G c c	15	71	14	74
C2	-172 to -153	t c C T C A a T T T C T GcT T T t T t	14	70	14	78
	+63 to +78	g a C C A T T C T G G T T G T t	13	68	13	76
C1	-112 to -94	t g C T C A a T A T T G T aa T C T t	13	68	13	76
	+71 to +90	t T C T g A T c A TgC T G G c T G T t	14	70	14	78

*Percentages expressed as number of bases correct over total number of bases in the consensus or in the particular sequence under examination, whichever is the greater.

†Number of correct bases if only the central 17 bp of the consensus sequence is considered.

progesterone receptor shows some specificity of binding to cloned fragments of C3(1) that include the 5' flanking DNA (Dr. W. Schrader, personal communication). However, the observation of such binding does not presuppose any functional significance. Within the MMTV genome there are several specific binding sites for the glucocorticoid receptor, but these are not all functionally equivalent. For example, the binding sites at -300 cannot functionally replace those binding sites between -100 and -200 because viral transcription becomes refractory to induction by dexamethasone if these 100 bp alone are deleted (Dr. G. Ringold, personal communication; Buetti and Digglemann, 1983; Hynes et al., 1983).

Nevertheless, it is interesting that the examination of the C3(1) 5' flanking DNA has revealed so many putative control sequences which all map between the Alu-like element and the cap sites, as predicted. Furthermore, for several (probably erroneous) reasons a region of 40 bp, centred at -100, may contain sequences important to the regulation of C3 expression (see Fig. 27). Obviously the next stage would be to assess if these sequences, pinpointed by structural similarities, actually fulfil any role in the androgen controlled expression of C3(1). One way of approaching this would be to perform functional tests along similar lines adopted for other steroid inducible genes. Thus 5' sequences must be progressively deleted and the residual activity of the gene remnant assayed within an androgen regulatable transcription assay. This may involve gene transfection into the androgen responsive S115 cell line (Page and Parker, 1983) or it may prove advantageous to develop chimeric C3 genes for analysis in short-term primary cultures of ventral prostate. The use of an homologous system is likely to optimize transcription and therefore facilitate what are technically difficult experiments. In this way it may be possible to determine if the removal of the sequences discussed above has an adverse effect on the ability to regulate C3(1) expression with androgens. Indeed it may be possible to physically and therefore functionally separate those elements allowing RNA polymerase II binding and detection of correct initiation sites from sequences involved in androgen inducibility. This has been found possible for the glucocorticoid regulation of MMTV expression (Chandler et al., 1983; Buetti and Digglemann, 1983), but must also be examined for the action of other steroids on their specific genes.

More importantly, in assessing the possible mechanisms whereby

androgens control gene expression, one must not become too confined within the Jensen model for steroid action as propounded in the Introduction. The best evidence that steroids can directly influence specific gene transcription at the level of the S-Rc complex binding to regulatory gene flanking sequences has come mainly from work on glucocorticoid regulated genes (see Table 2), especially the virus MMTV. However, this need not be the only level of action of a steroid hormone. Recent studies on the control of expression of the vitellogenin gene in the livers of male Xenopus have shown that oestrogen mediates at three levels: activation of vitellogenin gene transcription, induction of a 30-50 fold increase in the absolute rate of nuclear RNA synthesis and selective stabilization of vitellogenin mRNA in the cytoplasm (Brock and Shapiro, 1983). Moreover, it need not be the S-Rc complex itself that interacts with particular genes. It would not really alter the central tenets of Jensen's model if it were proposed that the S-Rc complex might interact with an 'inducer gene' so raising the levels of a transcript or protein which becomes the effector of steroid induced expression from a specific set of genes. In this regard it is worth recalling the work on the effect of ecdysone on the salivary gland of Drosophila larvae (Ashburner et al., 1973). As mentioned in the Introduction, some chromosomal puffs were induced very rapidly in the presence of steroid; a response which was not blocked by the addition of inhibitors of protein synthesis. However, a pattern of 'late puffs' was also observed, after a lag of at least three hours, but these puffs did not appear in the absence of cellular protein synthesis. The inference from these results is that a few chromosomal sites can interact with hormone and their activation leads to the induction of expression from other genes. Similar observations have also been made in mammalian systems. For example, of several marker proteins whose synthesis increases in response to dexamethasone in the rat hepatoma (HTC) cell line, at least one (α_1 acid glycoprotein) is not induced in the presence of cycloheximide (Feinberg et al., 1983).

Therefore, 'alternative' mechanisms of regulation should be considered to account for the androgen control of C3 gene expression, especially as the in vivo levels of C3 induction were not observed on transfection of the genes into S115 cells (Page and Parker, 1983). Moreover, preliminary studies attempting to monitor specific binding of the androgen receptor to cloned C3(1) DNA have proved unsuccessful (Dr. E. Mulder, personal communication). It must also be remembered that testosterone seems to act post-transcriptionally to promote steady-

state mRNA levels from all three prostatic steroid binding protein genes (Page and Parker, 1982). It is also intriguing that within the ventral prostates of juvenile rats (10-21 days old) mature levels of C2 mRNA have recently been recorded (Carter et al., 1983), but production of the C2 polypeptide is only at ~20% of normal levels during this stage of development (Heyns et al., 1978b; Higgins et al., 1982). Therefore, consideration of possible control mechanisms involving an interaction by androgens at the level of maturation of C3 mRNA, or its entry into polyribosomes, must not be overlooked.

To conclude, this project was begun in order to analyse a gene whose expression had been shown to be regulated by androgenic steroids. In the event, two highly homologous genes have been isolated and characterized, only one of which is thought to be expressed in vivo. The C3 genes therefore now provide an excellent model system to study not only the control of expression by androgens, but also more generally, why some genes are active while other, apparently similar, genes are silent within the same tissue. The answers to these questions are fundamental to an understanding of vertebrate development.

SECTION M: MATERIALS AND METHODS

M1. MATERIALS

1.1 Chemicals

Most commonly used chemicals were of analar grade and were purchased from British Drug Houses (B.D.H.), Fisons or Sigma. Components of bacterial broths were bought from Difco. Other chemicals, together with specific comments, were as follows:-

Agarose	Seakem, M.E. Miles.
α Amanitin	Boehringer. Dissolved to 2 mg/ml in 10% methanol. Stored -20°C.
Ampicillin	Beecham (Penbritin)
Bovine Serum Albumin (B.S.A.)	Sigma or Calbiochem (for use in enzyme buffers).
5 bromo, 4 chloro, 3 indoyl β galactoside (BCIG)	Sigma. Made up to 20 mg/ml in deionized formamide. Stored at 4°C.
Dextran Sulphate	Pharmacia.
Deionizing Resins	Amberlite MB-1 monobed resin (B.D.H.). AG 501-X8 20-50 mesh (Biorad).
Foetal Calf Serum	Gibco.
Ficoll 400	Pharmacia.
Formamide	Fisons. Deionized using AG 501-X8.
Isopropyl-β-thiogalacto-pyranoside (IPTG)	Sigma. Made up to 24 mg/ml in DW. Stored at 4°C.
Linkers	New England Biolabs.

β-mercaptoethanol	Koch-Light. Assumed 10 M concentration.
Nonidet P40 (NP40)	B.D.H.
Nucleic Acids	RNA and DNA used as carrier in precipitations and hybridizations. Sigma. DNA was boiled and sonicated to made short ss DNA.
Nucleotide Triphosphates	rNTP, dNTP, ddNTP. Sigma. 20 mM stocks in TE buffer, stored at -20°C.
Oligo dT Cellulose	Collaborative research.
Oligo $(dT)_{12-18}$	Miles.
Phenol	Fisons. Redistilled, buffered with TE (pH 8.0) and supplemented with 0.1% 8 hydroxyquinoline (Sigma).
Primer	Sequencing pentadecamer oligonucleotide. New England Biolabs.
Radioisotopes	Amersham International.
Sephadex	Pharmacia. G50 and G150.
Silane	Wacker Chemie, Munich (γ-methacryl-oxypropyltrimethoxysilane).
Siliconizing agent	Repelcote. Hopkins and Williams.
T.E.M.E.D.	Biorad.

1.2 Enzymes

Bal 31 nuclease	Bethesda Research Laboratories.
Calf Intestinal Phosphatase (CIP)	Boehringer. Made up in recommended dilution buffer at 50 U/ml. Stored at -20°C.

Lysozyme	Sigma. Made up fresh.
Pancreatic DNase 1	Worthington. Further purified to remove RNase, as described by Maxwell et al. (1977) by passing the DNase through a column of agarose 5' 4-amino-phenyl phosphoryluridine 2'(3') phosphate (Miles)equilibrated with 0.02 M NaOAc, pH 5.2. Stored in aliquots at -20°C.
Proteinase K	Merck.
RNase A	Worthington. Diluted 1 in 200 and heated 10 mins at 70°C before use.
Restriction Endonucleases	New England Biolabs.
Reverse Transcriptase	Life Sciences, Inc.
DNA Polymerase 1	Boehringer. Kornberg and Klenow enzymes.
Polynucleotide Kinase	P.L. Biochemicals. T_4 enzyme.
DNA Ligase	New England Biolabs. T_4 enzyme.

1.3 Other Materials

X-Ray film	Fuji X-Ray film. Kodak X-Omat AR film.
Screens	Dupont. Lightning Plus.
Plastikard	Slatersplastikard, England.
Centrifuge tubes	Corex. 30 ml and 15 ml for RNA and DNA preparations. Acid washed and siliconised. Sorvall 250 ml polypropylene bottles for plasmid and λ DNA preparations. Oakridge tubes for Beckman 50 Ti rotor.

Nitrocellulose	Schleicher and Schuell BA85 0.45μ for λ-screening. Sartorius 0.1μ for Southern Blotting. Millipore 90 mm circles, 0.45μ, boiled and dried before use in Grunstein and Hogness screening.

1.4 Standard Buffers and Culture Media

TBE	10 x buffer: 108 g/l Tris base, 55 g/l boric acid, 9.5 g/l EDTA, pH 8.3.
TE	10 mM Tris-Cl, pH 8.0, 1 mM EDTA. 10 x TE contained 10 mM EDTA.
NTE	10 mM Tris-Cl, pH 7.5, 0.1 M NaCl, 1 mM EDTA.
SSC	20 x buffer: 30 M NaCl, 0.3 M sodium citrate, pH 7.0.
SET	10 x buffer: 1.5 M NaCl, 0.3 M Tris-Cl, pH 8.0, 2 mM EDTA.
Denhardts	Denhardt (1966). 10 x buffer: 0.2% Ficoll, 0.2% BSA, 0.2% Polyvinylpyrollidine.
SSCP	2 x buffer: 0.24 M NaCl, 30 mM trisodium citrate, 26 mM KH_2PO_4.
Versene	0.54 mM EDTA in PBSA plus phenol red.
E4	Eagle's minimal medium.
Formamide dye	Deionized formamide with 0.2 M EDTA and xylene cyanol and bromophenol blue dyes.
Phenol mix	50:50:1 phenol:chloroform:isoamylalcohol, buffered with TE.

1.5 Bacterial Broths and Plates

L-Broth	5 g/l NaCl, 5 g/l yeast extract, 10 g/l tryptone, 1 g/l glucose, pH 7.0. For plates 11 g/l agar was added.
T-Broth	10 g/l tryptone, 5 g/l NaCl. T-broth + M had an additional 0.2% maltose.
NZY-Broth	5 g/l NaCl, 2 g/l $MgCl_2$. $7H_2O$, 10 g/l NZ amine A, 5 g/l yeast extract. For plates 11 g/l agar was added. For top agarose 6 g/l agarose was added.
2 x TY-Broth	16 g/l tryptone, 5 g/l NaCl, 10 g/l yeast extract.
H-Plates	12 g/l agar, 8 g/l NaCl, 10 g/l tryptone. H-Top agar had 8 g/l agar.
Minimal plates	2 g Agar in 80 ml DW. Added 20 ml sterile 5 x salts 100 µl sterile 20% $MgSO_4$ 1 ml sterile 20% glucose 5 x salts: 5.25% K_2HO_4 2.25% KH_2PO_4 0.5% $(NH_4)_2SO_4$ 0.25% Na citrate, $2H_2O$.

All were sterilized by autoclaving except the 5 x salts which were filter sterilized.

1.6 Cloning Hosts and Vectors

E.coli LE392	Prepared by L. Enquist from strain ED8654 (Murray *et al.*, 1977). Gift of A. Hayday. Grown in T-broth + M from single colonies streaked onto NZY plates.

E.coli HB101	Boyer and Rouland-Dussoix (1969). Grown in L-broth from single colonies streaked onto L-broth plates.
E.coli JM101/3	Messing (1979); Messing *et al.* (1981). Gift of S. Redmond. Maintained on minimal plates, restreaked every 6-8 weeks, to ensure retention of the F-episome. Grown from single colonies in 2 x TY broth.
pAT 153	Twigg and Sherratt (1980).
ds M13 mp 7, 8 and 9	Messing *et al.* (1981); Messing and Vieira (1982). M13 mp 7 gift of Dr. M. Jones. M13 mp 8 and 9 bought from New England Biolabs.

1.7 λ Recombinant DNA Libraries

Partial Eco RI and Hae III rat DNA libraries in λ Charon 4A were the generous gifts of Drs. T. Sargent, B. Wallace, L. Jagodzinski and J. Bonner, Caltech, Pasadena, U.S.A. A complete Eco RI library was constructed in λgtWES.λB by Roger White of our laboratory and kindly given to me to screen.

Phage titres:

p Eco RI	$3.0\ 10^8$ p.f.u./ml
p Hae III	$3.7\ 10^9$ p.f.u./ml with 0.09% contamination with pBR322.
c Eco RI	$6.0\ 10^6$ p.f.u./ml.

1.8 Cell Lines

The mouse mammary carcinoma cell line Shionogi 115 (S115) possesses androgen receptors (King *et al.*, 1976) and therefore is useful in the study of androgen controlled gene expression. A variety of S115 cell lines containing stably integrated copies of C3 rat genes were kindly provided, from frozen stocks, by Dr. M. Page. These cell

lines had been obtained using the dominant acting selection vector pSV2-gpt to introduce particular restriction fragments from C3(1) or C3(2) into S115 cells. The final C3 gene content of each cell line used was as follows (Page and Parker, 1983; M. Page, unpublished):

CELL LINE	GENE	RESTRICTION FRAGMENT	GENE COPIES/ CELL
A1 cells	C3(2)	9kb Bam HI	50-100
50 cells	C3(1)	9kb Bam HI	50
C5 cells	C3(1)	12.5kb Eco RI	1-3
SV2 and SV6 cells	C3(1)	5.8kb XbaI/Bam HI Possessed only 800bp of 5' flanking DNA	unknown

These cell lines were originally selected by growth in HAT medium supplemented by mycophenolic acid and xanthine (Page and Parker, 1983). However, cultures grown up from frozen stocks were maintained in standard S115 medium of E4 plus 2% foetal calf serum, 30 mM unbuffered HEPES and 3×10^{-8} M testosterone.

M2. METHODS

2.1 General Manipulations of DNA

The conditions for general enzymic treatments of DNA used in this thesis are detailed here. Between enzyme treatments the DNA samples were generally deproteinized by phenol extraction and then precipitated by addition of salt (either NaCl to 0.1 M or NaOAc, pH 7.0 to 0.3 M) and 2 volumes of ethanol. DNA in large volumes was recovered by the addition of 1 to 1½ volumes isopropanol.

a) Restriction endonuclease digestions.

DNA digestions were performed in a standard buffer of 50 mM Tris-Cl, pH 7.5, 10 mM $MgCl_2$, 10 mM mercaptoethanol, which was supplemented with 50-150 mM NaCl for certain restriction enzymes, as recommended by Biolabs. Digestions of cloned DNA were 1-3 hrs in 10-20 µl mixes, while cell DNA was digested in 50-75 µl for 20-30 hrs with 2-3 additions of enzyme. The temperature of these incubations was generally 37°C, but certain restriction enzymes, such as $BstE_{II}$, required higher temperatures (60°C).

b) Removal of 5' phosphate groups (Maxam and Gilbert, 1980).

Calf intestinal phosphatase (CIP) was used to generate 5' hydroxyl groups on restricted DNA samples. DNA (1-5 µg) in a 30 µl mix containing 10 mM Tris-Cl, pH 9.5, 1 mM spermidine, 0.1 mM EDTA and 3 µl CIP (50 U/ml) was incubated for 30 mins at 37°C. The CIP enzyme was then inactivated at 75°C for 1 hr.

c) Ligation of DNA.

DNA ligations were carried out in 10 µl mixes with approximately 10 ng vector and either an equimolar amount of genomic DNA fragment or 100-fold excess of linkers. The ligation mix also contained 1 mM ATP, 10 mM $MgCl_2$, 5 mM DTT, 50 mM Tris-Cl, pH 7.5 and 100-500 U T_4 DNA ligase. Incubations were at 4°C o/n. Blunt end ligations and those using linkers required the higher amounts of ligase and incubation at 14°C o/n.

Vector DNA was generally treated with CIP prior to ligation to reduce the recircularization efficiency and increase the likelihood of generating recombinant DNA molecules.

d) 'Blunt-ending' DNA.

The 5' overhangs created by some restriction endonuclease digestions could be filled in by treating DNA (1-5 µg) with 1 U Klenow

DNA polymerase. Incubations were for 10 mins at 37°C in 50 µl of a buffer containing 50 mM Tris-Cl, pH 7.4, 10 mM $MgCl_2$, 10 mM DTT and each dNTP at 1 mM.

e) Resecting DNA.

Bal 31 nuclease was used to digest back DNA from a unique restriction site to generate smaller (deletion mutant) cloned inserts. Restricted DNA (10 µg) was resuspended in 200 µl of a buffer containing 200 mM NaCl, 12 mM $CaCl_2$, 12 mM $MgCl_2$, 20 mM Tris-Cl, pH 8.0, 1 mM EDTA and 1 U Bal 31 nuclease. Incubation was at 30°C for a total of 10 mins, but 40 µl aliquots were withdrawn into phenol at 2 min intervals, extracted and precipitated. Half of the DNA recovered from each aliquot was digested with a second enzyme and run on an agarose gel to determine the extent of the resection. Under these conditions 100 to 150bp were removed per min.

2.2 Preparation of RNA Samples

During the isolation and handling of RNA samples, sterilized and siliconized glass- and plastic-ware was used throughout.

a) Isolation of total and p(A^+) RNA from rat ventral prostate (Parker and Mainwaring, 1977).

Ventral prostates were cleaned of connective tissue and frozen on dry ice. Tissue (2.5 g) was homogenized with 8 ml phenol and 8 ml aqueous buffer (0.5% SDS, 100 mM NaCl, 100 mM EDTA, pH 7.5). The homogenate was then made up to 25 ml each of phenol and aqueous buffer and shaken 15 mins at RT before centrifugation to separate the phases. The aqueous layer was extracted further with 50:50 phenol:chloroform and chloroform and then set aside. Each extraction left a fluffy white interface between the two phases. This was removed, supplemented with 5 ml aqueous buffer and further extracted in three stages as above. The final aqueous layer was pooled with that reserved from the main extraction and the combined sample was re-extracted with phenol:chloroform and chloroform. Total nucleic acid ($\sim\frac{1}{3}$ DNA, $\sim\frac{2}{3}$ RNA) was recovered by addition of NaCl to 0.1 M and precipitation o/n at -20°C with 2½ volumes of ethanol. The centrifuged pellet was washed (70% ethanol), dried and resuspended in 4-5 ml DW on ice.

To prepare total prostate RNA, total nucleic acid (400-500 µg) was treated with 20 µg/ml RNase free DNase 1 for 30 mins at 37°C in a 250 µl mix containing 10 mM $MgCl_2$ and 50 mM Tris-Cl, pH 7.6. The

sample was then phenol:chloroform extracted twice and the RNA recovered by precipitation, resuspended in TE buffer and stored at -70°C. The integrity of the RNA was checked by observing the running of rRNA bands on a formaldehyde gel, compared to an untreated sample. The presence of tight bands indicated that little RNA degradation had occurred during DNA removal.

To prepare total p(A^+) RNA, total nucleic acid was made 0.5 M KCl and 0.01 M Tris-Cl, pH 7.5 and allowed to drip slowly through a 1 cm oligo dT cellulose column, equilibrated with the same buffer. The eluate was collected and reapplied. The column was then extensively washed (30-50 ml) with KCl/Tris buffer until the eluate gave an OD_{260} reading of zero. Total p(A^+) RNA was eluted with 10 mM Tris-Cl, pH 7.5. After collecting 6 to 8 1 ml fractions, those containing RNA were pooled and the RNA precipitated o/n. The washed and dried pellet was resuspended in 500 µl TE and stored at -70°C. Usually 400-500 µg p(A^+) RNA were recovered from 2.5 g ventral prostate tissue.

b) Isolation of cytoplasmic RNA from cultured cells.

Log phase cells (~90% confluent) on a 15 cm culture dish were removed using trypsin in versene, washed in PBSA and repelleted. The cells were resuspended in 2.5 ml NTE buffer and left 15 mins on ice. Then NP40 was added to 0.5% and the cells lysed during a further 10 min incubation on ice. Nuclei were removed by centrifugation and the supernatant extracted twice with phenol mix and once with chloroform. RNA was recovered from the final aqueous phase by precipitation. A further precipitation was usually required to concentrate the RNA sufficiently for primer extension analysis.

Yields of cytoplasmic RNA varied greatly, but were usually in the range 100-350 µg RNA per 15 cm dish.

2.3 Gel electrophoresis

a) Neutral agarose gels (McDonell et al., 1977).

These were used to analyse restricted ds DNA samples and to check the purity of isolated ss M13 recombinants. Agarose was dissolved by boiling in 1 x TBE at 0.5 to 2.5% and cast on a glass plate 6.0 cm or 14.0 cm wide by 19.0 cm long. Sample slots were formed using a plastic 'comb', to give a horizontal gel.

DNA samples were loaded with the addition of $\frac{1}{3}$ volume sample buffer (50% glycerol, 1 x TBE, 2.5% SDS, 125 mM EDTA, bromophenol blue).

Restricted λ-DNA samples were heated 5 mins at 65°C with 0.5% SDS and 1 mM EDTA before addition of sample buffer. The gels were run in 1 x TBE at 30V o/n or 100V for 2-3 hrs. A drop of 10 mg/ml ethidium bromide added to the running buffer allowed the DNA bands to be observed and photographed under U.V. light.

b) Alkaline agarose gels (McDonell et al., 1977).

The ss DNA samples from some S1 mapping experiments were analysed on these gels. Agarose gels were cast as above, but gel buffer and running buffer consisted of 30 mM NaOH and 2 mM EDTA. Samples were first denatured for 10 mins at RT by the addition of NaOH to 0.2 M and EDTA to 20 mM, then loaded with ¼ volume sample buffer (20% sucrose, 0.05% bromophenol blue) and run at 250 mA for 3-4 hrs.

c) Formaldehyde agarose gels.

These were used to examine RNA samples. Agarose (1.5 g) was dissolved in 72 ml DW plus 10 ml of 10 x running buffer (200 mM MOPS, 10 mM EDTA, 50 mM NaOAc, pH 7.0). After cooling, 18 ml formaldehyde (38% stock) was added and the gel cast as above. Samples in 5 µl DW were made up to 20 µl with sample buffer (200 µl 10 x running buffer, 1 ml formamide, 356 µl formaldehyde) and heated 5 mins at 60°C. Finally, 8 µl dye (7.5% Ficoll, bromophenol blue) were added. Gels were run in 1 x running buffer at constant current:20 mA o/n or 90 mA for 3 hrs. To detect the RNA samples, the gel was soaked 30 mins in 1 x running buffer plus a drop of ethidium bromide, destained o/n in 1 x running buffer and photographed in U.V. light.

d) Acrylamide sequencing gels (Sanger and Coulson, 1978).

Long, thin acrylamide gels were used to run out the ss DNA products from sequencing reactions and primer extension or S1 mapping experiments. Stock acrylamide contained 6% acrylamide (19:1 acrylamide to bisacrylamide) and 7 M urea in 1 x TBE. This was prepared by dissolving together the urea and acrylamide and deionizing this solution with Amberlite resin before addition of the TBE buffer. The stock was stored at 4°C for up to 1 month.

The gel plates were 23 cm wide by 43 cm long. The inner face of the front, notched plate was siliconized to encourage the acrylamide to adhere to the backplate when the gel was finally dismantled. However, the gels were even more convenient to handle if the acrylamide was covalently attached to the backplate (Garoff and Ansorge, 1981). The inner face of the backplate was evenly spread with a mixture

of 0.15 ml glacial acetic acid dissolved in 5 ml DW and 0.15 ml silane dissolved in ethanol. This mixture was allowed to dry on the glass and the surface was then washed with ethanol.

The treated glass plates were taped together, separated by spacers of 0.3 mM Plastikard. Each gel was poured with 50 ml stock acrylamide plus 35 µl TEMED and 250 µl 10% ammonium persulphate. Slots for loading samples were formed using a template cut from Plastikard to give tracks 6 mm wide and 8 mm deep. Gels were generally poured a day in advance and stored flat at 4°C.

Before loading, the gels were pre-run for 15 to 30 mins and the wells extensively washed out to remove urea. Samples were loaded using drawn out melting point tubes and the gels were run at 33 mA, 50 W (1400-1600V) in 1 x TBE.

e) Small acrylamide gels.

Gels of ~15 ml volume, 16 cm wide by 13 cm long and 1.0 mm thick (Teflon spacers) were poured using the stock acrylamide. These were used for separating ds DNA fragments upto 1.0kb long. Gels were run at 150V for 2-3 hrs. Labelled fragments were located by direct autoradiography. Unlabelled DNA fragments were detected by soaking the gel in 1 x TBE containing a drop of ethidium bromide for 30 mins and viewing under U.V. light.

DNA fragments were eluted from acrylamide (Maxam and Gilbert, 1980) by mincing the slice containing the fragment and incubating it o/n at 37°C in 400 µl of buffer (500 mM NH_4OAc, 1 mM EDTA, 0.1% SDS). The acrylamide was filtered off using Whatman GF/C filters, carrier tRNA added to 10 µg/ml and the DNA precipitated by adding an equal volume of isopropanol. The precipitated DNA was redissolved in 0.3 M NaOAc and reprecipitated with ethanol, washed, dried and resuspended in 10-50 µl TE. Recovery of fragments up to 1.0kb was 50-80% efficient. Larger fragments (up to 3.0kb) were more difficult, but not impossible to isolate this way. Incubation o/n at 45-55°C usually improved recovery.

2.4 $[^{32}P]$-labelling

Nucleic acids were labelled using high specific activity nucleotides, either $[\alpha^{32}P]$-dNTP (>3000 Ci/mmol) or $[\gamma^{32}P]$-ATP (5000 Ci/mol). Particular 5' or 3' end-labelled DNA fragments were recovered from sequencing (ss fragments) or small (ds fragments) acrylamide gels. Otherwise labelled nucleic acids were separated from unincorporated radioisotope by passage down a sephadex G50 column, equilibrated with 0.1% SDS, 100 mM NaCl and 10 mM Tris-Cl, pH 7.5. The activity of all labelled samples was estimated by liquid scintillation of a small aliquot of the recovered nucleic acid.

a) 5' end-labelling (Maxam and Gilbert, 1980).

Polynucleotide kinase was used to label overhanging 5' ends of suitably restricted, cloned DNA. Digested DNA (5-10 µg) was treated with CIP, as detailed above. The DNA was then incubated 15 mins at 37°C with 5-10 U kinase and 100 µCi $[\gamma^{32}P]$-ATP in 50 µl of a buffer containing 10 mM $MgCl_2$, 5 mM DTT, 0.5 mM spermidine-HCl, 0.05 mM EDTA and 50 mM Tris-Cl, pH 9.5.

RNA samples could also be labelled this way if 5' hydroxyl groups were first generated by mild alkali treatment (Maizels, 1976). Total RNA (1 µg) was incubated in 20 µl 50 mM Tris-Cl, pH 9.5 for 5-15 mins at 65°C and then kinased as above.

b) 3' end-labelling.

Cloned DNA, digested to give a 5' overhang, was 3' end-labelled by filling in the overhang using Klenow DNA polymerase and 100 µCi of one or more $[\alpha^{32}P]$-dNTP's, as described for the 'blunt-ending' of DNA.

To increase the incorporation of labelled nucleotides, a modification of this method (McKnight et al., 1981) was also used. The 5' DNA overhang was extended by incubating the DNA with 100 U/ml exonuclease III in 50 µl of a buffer containing 60 mM NaCl, 5 mM $MgCl_2$ and 1 mM DTT for 2 mins at 10°C. The reaction was stopped by the addition of SDS to 1% and EDTA to 10 mM, phenol extracted and the DNA precipitated. The extended overhang was then filled in as before using one or two appropriate labelled nucleotides.

c) Nick-translating DNA (Rigby et al., 1977).

Nick-translations were performed in 25 µl mixes containing 0.1 to 0.5 µg DNA, 25 µCi $[\alpha^{32}P]$-dCTP, 50 mM Tris-Cl, pH 8.0, 5 mM $MgCl_2$, 10 mM mercaptoethanol, 2 µM each of dATP, dGTP, dTTP, 40 µg/ml BSA and 5 U DNA polymerase 1. The polymerase preparation was contaminated

with some DNase, but to ensure that sufficient 'nicks' were produced for good labelling, extra DNase 1 was added to 20 ng/ml. Incubations were left 3 mins at RT to allow the DNase to act and then maintained at 14°C for 3 hrs. The labelled DNA was finally denatured for 10 mins at RT by addition of NaOH to 0.4 M.

d) Preparation of cDNA.

cDNA was prepared in 25 µl mixes containing 1-2 µg p(A^+) RNA, 25 µCi [α^{32}P]-dCTP, 60 mM KCl, 10 mM DTT, 6 mM $MgCl_2$, 50 mM Tris-Cl pH 8.3, 50 µg/ml oligo$(dT)_{12-18}$, 1 mM each of dATP, dGTP, dTTP and 10 U reverse transcriptase. Incubation was for 1 hr at 37°C followed by RNA hydrolysis for 30 mins at 37°C on addition of NaOH to 0.4 M.

2.5 Isolation of Genomic Clones from λ DNA Libraries

The three libraries screened were all grown in E.coli LE392. Bacterial cultures were grown for 3-4 hrs in T-broth + M from single colonies streaked onto NZY-plates. The cultures were pelleted and resuspended in 0.4 x culture volume 10 mM $MgSO_4$ to give an OD_{600} of 1-2. These bacteria were then ready for phage infection.

a) Plating bacteriophage.

Primary screening was performed on large plates (22 x 22 cm) so as to give ~95% confluent lysis corresponding to ~70,000 plaques/large plate. Subsequent screens used 90 cm petri dishes at 400-1000 plaques/dish (5-10% confluent lysis).

Plating conditions were as follows for the large plates with volumes for the petri dishes given in parentheses. Suitable dilutions of the λ-library in 1 ml (200 µl) phage buffer (10 mM Tris-Cl pH 7.5, 5 g/l NaCl, 1 mM $MgSO_4$, 0.01% gelatin) were absorbed onto 1 ml (200 µl) LE392 for 20-30 mins at 37°C. Then 25 ml (2.5 ml) molten NZY-agarose at ~50°C was added to each incubation, mixed and spread over a well-dried plate. When set, the plates were inverted and maintained at 37°C o/n.

b) Screening (Benton and Davis, 1977; Maniatis et al., 1978).

The bacterial plates were placed for 1-2 hrs at 4°C to harden the top-agarose. A sheet of nitrocellulose was carefully lowered onto the plate for 1-2 mins and orientated with pin pricks through the filter and agar. The filter was then removed and the DNA fixed by washing in 100 mM NaOH, 1.5 M NaCl followed by neutralization in 200 mM Tris-Cl, pH 7.5, 2 x SCCP. Meanwhile a duplicate lift was

taken from the plate. Dried filters were baked for 2 hrs in an 80°C vacuum oven before hybridization in sealed plastic bags (4 lifts/bag) as follows: i) 3 hrs at 65°C with 100 ml 4 x SET, 10 x Denhardts, 0.1% SDS; ii) 1 hr at 65°C with 100 ml buffer (i) supplemented with 50 µg/ml ss DNA, 10 µg/ml polyA and 0.1% NaPPi; iii) o/n at 65°C with 20 ml buffer (ii) containing 10^6 cpm per filter of radiolabelled DNA probe; iv) 1 hr at 65°C with 100 ml buffer (i) supplemented with 10 µg/ml polyA and 0.1% NaPPi. The lifts were then extensively washed in a plastic tray in 250 ml of successive buffers: i) 3 x 30 mins at 65°C with 3 x SET, 0.1% SDS, 0.1% NaPPi; ii) 2 x 30 mins at 65°C with 1 x SET, 0.1% SDS, 0.1% NaPPi; iii) 30 mins at RT with 4 x SET.

The filters were dried and autoradiographed with preflashed film (Laskey and Mills, 1977) for 24-60 hrs at -70°C with screens. Spots on the film due to the presence of λ-clones can be indistinguishable from the small amount of background but the use of duplicate lifts usually enabled true λ-clones to be unambiguously identified.

c) Isolation of recombinant clones.

Agar plugs were withdrawn from regions of the bacterial plates identified (by screening) as containing recombinant clones. Phage particles were eluted, in 1 ml phage buffer, and titred (5×10^7 to 5×10^8 p.f.u./ml). The phage were then replated and rescreened until single, pure recombinant clones were recovered. Four rounds of screening were always necessary even though apparently pure, single colonies could be picked from 3rd screen plates.

Plate lysate stocks of pure clones were made by overlaying a confluently lysed petri dish with 4 ml phage buffer for 4-24 hrs at 4°C. The buffer was recovered, titred (10^9-2×10^{11} p.f.u./ml) and stored at 4°C.

d) Preparation of λ recombinant DNA.

Phage particles (10^7 p.f.u. λgtWES.λB or 2×10^8 p.f.u. λ Charon 4A) were incubated with LE392 (10^{10} cells) in 300 µl phage buffer for 10 mins at 37°C. The incubation was then used to innoculate 700 ml NZY-broth, supplemented with 0.1% casamino acids and prewarmed to 37°C. Bacterial cultures were shaken o/n at 37°C and shaken a further 10 mins next morning after addition of 5 ml chloroform and 56 g NaCl. Bacterial debris were removed by centrifugation. The supernatant was decanted into a fresh flask and supplemented with $MgSO_4$ to 10 mM and PEG 6000 to 10% w/v. Phage particles were precipi-

tated by slowly stirring at 4°C for at least 1 hr to dissolve the PEG. Precipitated phage were recovered by centrifugation (30 mins at 7000 r.p.m., 4°C) and the pellet resuspended in 10 ml 50 mM Tris pH 7.5, 10 mM $MgSO_4$ with 7.5 g CsCl. This solution was spun at 40,000 r.p.m. in an Oakridge tube at 20°C for 20-24 hrs. Under these conditions phage particles sediment ~¾ down the tube and can be observed as a bluish line in strong light. The phage particles were removed in as small a volume as possible and dialysed o/n against two changes of TE. Phage DNA was recovered on gentle extraction, twice with phenol and once with chloroform and was further dialysed against TE. The final DNA sample was stored at 4°C. Yields of phage DNA were usually very low, 100-500 µg DNA.

2.6 Subcloning into pAT 153 and M13

To facilitate the restriction enzyme mapping and sequencing of the genomic λ-clones, they were subcloned, wholly or partially, into plasmid pAT 153 and ds forms of M13 bacteriophage mp7, 8 and 9. Although these two vectors have quite different replicative processes, they both possess a circular genome. Thus, the methods used in their manipulation to generate recombinant clones are very similar.

a) Generation of competent bacteria.

To enable bacteria to take up *in vitro* modified plasmid or M13, their walls were made leaky using $CaCl_2$. Small o/n cultures of *E.coli* HB101 (for pAT) or *E.coli* JM101 (for M13) were diluted 1 in 50 into 50 ml L-Broth or 2 x TY, respectively, and shaken 1-3 hrs at 37°C to give OD_{550} of 0.3. Bacterial cells were then pelleted, resuspended in 25 ml sterile 50 mM $CaCl_2$ and incubated on ice for 20 mins. The competent cells were repelleted, resuspended in 5 ml 50 mM $CaCl_2$ and stored on ice for up to 24 hrs. HB101 competent cells could be frozen in liquid nitrogen in 80 mM $CaCl_2$, remaining competent on thawing.

b) Generation of recombinant clones.

Genomic λ clones or large plasmid subclones (~2 µg) were digested with restriction endonuclease(s). Circular vector DNA (~2 µg) was also suitably restricted and treated with CIP, as above. The vector and genomic DNA were then ligated together in 10 µl mixes o/n as detailed before. This ligated DNA was used to transform the appropriate competent bacteria, 200 µl of which being added to each ligation reaction, mixed and incubated for 20-40 mins on ice. These incubations were heat shocked at 37°C for 3-5 mins. The two types of transformation

(plasmid and M13) were thereafter treated differently.

For HB101 cells, L-Broth was added to 900 µl and incubation at 37°C continued for a further 40-60 mins. Each transformation was then divided between 3 well-dried L-Broth plates, supplemented with either 15 µg/ml tetracycline or 100 µg/ml ampicillin to kill all bacteria that did not take up plasmid.

For JM101 cells, each transformation was mixed with 3 ml molten H-top agar (at ~50°C), 25 µl BCIG (20 mg/ml), 20 µl IPTG (24 mg/ml) and 200 µl fresh bacteria (o/n culture JM101 diluted 1 in 10 in 2 x TY and shaken 1 hr at 37°C). Each mix was tipped onto one well-dried H-plate.

All plates were allowed to dry, inverted and maintained o/n at 37°C. Identification and growth of plasmid and M13 recombinant clones also followed differing procedures.

c) Plasmid screening (Grunstein and Hogness, 1975).

To identify which bacterial colony carried the desired recombinant clone they were all transferred to nitrocellulose filters for *in situ* hybridization. Filters were left in contact with the plates for 1 min then removed, 'colony-side' up to 3 mm Whatman paper soaked in 0.5 M NaOH for 10 mins. The 'fixed' filters were then neutralized successively on Whatmann paper soaked in 1 M Tris, pH 7.5 (1-2 mins) and 1.5 M NaCl, 0.5 M Tris-Cl, pH 7.5 (5 mins). The bacterial plates were returned to 37°C for 2-3 hrs, to allow bacterial regrowth, then stored at 4°C.

Neutralized filters were dried, baked 2 hrs in an 80°C vacuum oven and prehybridized for 3 hrs at 65°C in 250 ml 5 x Denhardts. After a brief wash in 100 ml of hybridization buffer (5 x SET, 1 x Denhardts, 0.5% SDS), the filters were hybridized o/n in 20 ml containing ~5 x 10^6 c.p.m. radiolabelled DNA probe at 65°C. The filters were washed 2 x 2 hrs at 65°C with 250 ml 2 x SET, 0.5% SDS and for 3 hrs at RT with 250 ml 3 mM unbuffered Tris. At no time were the filters shaken in case the bound colonies were knocked off.

Positive colonies were located by o/n autoradiography of the dried filters at -70°C with screens and preflashed film (Laskey and Mills, 1977).

d) Plasmid Mini preps (Holmes and Quigley, 1981).

The plasmid content of positively hybridizing bacterial colonies was examined by preparing limited amounts of DNA from small cultures.

Single bacterial colonies were inoculated into 5 ml L-Broth plus appropriate antibiotic and shaken o/n at 37°C. An aliquot (1.5 ml) of each culture was spun 10 seconds in a microfuge and all the supernatant removed. The cell pellet was thoroughly resuspended in 350 µl Triton buffer (8% sucrose, 0.5% Triton X-100, 50 mM EDTA, 10 mM Tris-Cl, pH 8.0). To each tube was added 25 µl of a fresh solution of lysozyme at 10 mg/ml in 10 mM Tris-Cl, pH 8.0. Completed incubations were vortexed 3 seconds, immersed in boiling water for 40-60 seconds, then spun in a microfuge for 10 mins. Gelatinous pellets of bacterial DNA were removed using a cocktail stick and plasmid DNA was precipitated on dry ice by addition of 40 µl 3.0 M NaOAc, pH 7.0 and 450 µl isopropanol to the supernatant. Some 3-5 µg plasmid DNA were recovered and analysed by restriction enzyme digestion, though RNA in the samples had to be removed using RNase A. However, the plasmid DNA digested satisfactorily and could be directly recloned, into M13 vectors for example.

e) Large scale plasmid DNA preparation.

Desired recombinant clones were grown up in bulk culture and the DNA isolated using the rapid alkaline lysis procedure of Birnboim and Doly (1979), as modified by Ish-Horowicz and Burke (1981).

Bulk cultures (400 ml) of single bacterial colonies were shaken o/n at 37°C, then pelleted (5 mins at 4000 r.p.m.). Bacterial cells were resuspended in 20 ml lysozyme buffer (50 mM glucose, 25 mM Tris-Cl, pH 8.0, 10 mM EDTA, 5 mg/ml lysozyme) and incubated 10 mins at RT. This was followed by the addition of 40 ml 0.2 M NaOH, 1% SDS and incubation on ice for 5 mins. Bacterial DNA was precipitated for 15 mins on ice after addition of 20 ml 5 M KOAc, pH 4.8 (3 M KOAC plus 2 M glacial acetic acid). The precipitate was pelleted (10 mins at 5000 r.p.m.), the supernatant filtered through medical gauze and plasmid DNA recovered by addition of 0.6 volume isopropanol (~50 ml). The washed and dried plasmid DNA pellet was resuspended in 7.5 ml 10 x TE and neutralized with 100 µl 2 M Tris. To this solution was added 8.1 g CsCl, 0.5 ml 10 mg/ml ethidium bromide and the completed mix was spun in an Oakridge tube at 20°C, 40,000 r.p.m. for at least 40 hrs. Under these conditions plasmid DNA sediments lower in the gradient than residual bacterial DNA and was located under U.V. light and withdrawn in ~1 ml. Ethidium bromide was removed in 3-4 isobutanol extractions and plasmid DNA was recovered by precipitation (350-500 µg). Further purification from contaminating oligonucleotides was achieved by

passing the DNA sample down a sephadex G150 column, buffered with 0.1 M NaCl, 10 mM Tris-Cl, pH 7.5. The final DNA sample was stored at -20°C in TE buffer.

f) Preparation of ss M13 DNA.

White plaques, denoting recombinant subclones, were inoculated separately into disposable, capped plastic tubes containing 1.5 ml of a 1 in 20 dilution of an o/n culture of JM101 in 2 x TY. The tubes were shaken vigorously for 5-6 hrs at 37°C before pelleting the bacterial cells in microfuge tubes. Phage particles were precipitated from 1 ml of the supernatant, by addition of 250 µl 20% PEG 6000, 2.5 M NaCl, for 20 mins at RT. The particles were pelleted in a microfuge, all the supernatant was removed and the phage were resuspended in 100 µl TE. Deproteinization, using freshly buffered redistilled phenol (containing no hydroxyquioniline and stored at -20°C), released ss phage DNA which was precipitated, resuspended in 30 µl TE and stored at -20°C. Recovery and purity of the phage DNA was checked by running $\frac{1}{10}$th sample on a 0.8% neutral agarose gel. DNA sufficient for 5-6 dideoxy sequencing reactions was isolated in this way, but more could be acquired by transfecting an aliquot back into competent JM101 cells and repeating the purification procedures.

2.7 Analysis of Recombinant Clones

a) Southern blotting (Southern, 1975; 1980).

Cloned DNA (1 µg) was digested with restriction enzymes and samples run on neutral agarose gels in preparation for blotting. Gels were soaked in 0.5 M NaOH, 1 M NaCl for 1-2 hrs, to denature ds DNA, and neutralized for a further 2 hrs in 2.5 M NaCl, 0.5 M Tris-Cl, pH 7.5. (Alkaline gels were neutralized only; gels containing M13 clones were blotted directly).

DNA was transferred from the gels to nitrocellulose filters o/n, as detailed by Southern (1980), using 20 x SSC buffer. The filter was rinsed in 2 x SSC, dried and baked 2-4 hrs in an 80°C vacuum oven. These blots were generally hybridized using the formamide method, described below, but occasionally other methods were used as described in individual figure legends.

Nitrocellulose filters were prehybridized for 1 hr at 37°C with 20 ml buffer (3 x SSC, 2 x Denhardts, 50 µg/ml carrier RNA, 10 µg/ml carrier ssDNA, 50% formamide) in separate, sealed plastic bags. This was then replaced with 10 ml buffer supplemented with 10^5-10^6 c.p.m.

radiolabelled probe for o/n hybridization at 37°C. The blots were then washed in a plastic tray with 250 ml of successive buffers as follows: i) 20 mins at RT with 2 x SSC; ii) 3 x 30 mins at 50°C with 0.1 x SSC, 0.1% SDS; iii) 10 mins at RT in 0.1 x SSC. Autoradiography of the dried filters o/n with screens at -70°C followed; preflashing the film was not usually necessary.

b) R-looping

Genomic λ-cloned DNA (0.5 µg) was precipitated with rat ventral prostate p(A^+) RNA (1.0 µg) using salt and ethanol. The washed and dried samples were then resuspended in 20 µl of formamide buffer for hybridization. Various hybridization conditions were tried (varying the temperature and percentage of formamide); the most successful used a buffer of 70% deionized formamide, 400 mM NaCl, 100 mM PIPES pH 7.2 at 47.5°C. Samples were then diluted, spread on DW as described by Wahli et al. (1978) and the nucleic acids were prepared for electron microscopy, using platinum shaddowing, by Mrs. Rita Tilley.

R-loop sizes were estimated from measurements taken from photographs using a Hewlett-Packard Digitiser.

2.8 DNA sequencing

a) Dideoxysequencing

The methods described in this section, and those used for the generation of M13 subclones, were based on protocols kindly provided by Dr. F. Sanger's laboratory in Cambridge.

Generally four recombinant M13 clones were sequenced at a time. For each clone 6 µl (~0.5 µg) ss DNA, 2 µl sequencing primer (37.5 pmoles/ml) and 2 µl hybridization buffer (0.1 M Tris-Cl pH 7.4, 0.1 M $MgCl_2$, 0.5 M NaCl, 0.1 M DTT) were sealed into a 50 µl capillary. Each capillary was then placed in a small test tube, full of water, standing in an 85°C water bath. After 3 mins the complete test tube rack was withdrawn and allowed to cool to RT (20-30 mins). The capillary contents were evacuated into a microfuge tube and 25 µCi (8.3 pmoles) [α-^{32}P] -dCTP (3000 Ci/mmol) added. This sample was divided equally between four fresh tubes labelled A, G, C and T to which were added 2 µl of an appropriate nucleotide mix, detailed in Table 4, and 0.5 µl Klenow DNA polymerase (5 U/µl).

Table 4. Contents of 'dideoxysequencing' nucleotide mixes

MIX	µl 0.5 mM dNTP da	dG	dT	ddNTP	µl TE	Final [ddN] : [dN] ratio
A	2	40	40	4 µl 10 mM ddA	78	40:1
G	40	2	40	4 µl 10 mM ddG	78	40:1
C	30	30	30	6.8 µl 0.5 mM ddC	83	18:1
T	40	40	2	4 µl 10 mM ddT	78	40:1

All of these additions were made as separate drops on the sides of the tubes so the extension reactions could be started simultaneously by spinning the tubes briefly. Incubation at RT for 15 mins was followed by the addition of 2 µl chase (each dNTP at 0.5 mM in TE) to each tube for a further 15 mins at RT. The reactions were terminated with 4 µl formamide dye. Finally the samples were boiled for 5-10 mins and loaded onto 6% sequencing gels, each sample being divided between two gels. One gel was run for 1½-2 hrs; the other for 3½-4 hrs. These gels were either autoradiographed directly at -70°C or, to increase resolution, were first fixed in 10% glacial acetic acid and dried extensively with Kleenex tissues before autoradiography at RT. Exposures were for 2-16 hrs, without screens.

For cold prechase sequencing the ss clone and primer DNA's were hybridized as normal and transferred to a microfuge tube. Diluted chase (4 µl of a 1 in 750 dilution) together with 2 µl Klenow DNA polymerase were then added and the primer was extended using the cold dNTP's for 2-4 mins while the sample was divided between 4 microfuge tubes as before. Then 0.6 µCi [α^{32}P]-dCTP was added to each tube plus 2 µl of appropriate nucleotide mix (Table 4). The procedure was then as described above, except that samples were loaded onto one sequencing gel and run for 5-6 hrs. To increase resolution, these gels were not fixed to the glass back plate thus allowing the acrylamide to be transferred to Whatman 3 mm paper and dried down for direct autoradiography.

b) Chemical degradation sequencing (Maxam and Gilbert, 1980).

Cloned plasmid DNA (5 µg) was completely digested with a 5' overhang producing restriction enzyme and 5' end-labelled as previously described. The labelled DNA was cut asymmetrically with a second restriction enzyme and the required fragment isolated from a small acrylamide gel. Chemical degradation reactions were carried out on the recovered DNA as explicitly described by Maxam and Gilbert (1980) and the products run on a 6% sequencing gel. The gel was autoradiographed directly at -70°C, with a screen, for 48-72 hrs.

c) Handling of DNA sequences.

All sequence information read from the gels was typed into computer storage programmes which were used to overlap and order the sequences with the DBUTIL programme (Staden, 1980). The final sequences obtained were examined for restriction enzyme sites, protein coding regions, symmetries and homologies using the SEQ programme (Brutlag et al., 1982).

2.9 Transcript Mapping.

The 5' and 3' ends of RNA transcripts were mapped to the encoding DNA using S1 mapping or primer extension techniques with end-labelled probes or primers. Structures of individual probes and primers are given in the figure legends while end-labelling methods are described in section 2.4 above.

a) Primer extension (Proudfoot et al., 1980).

This method of 5' mapping required the extension of a short ss DNA primer, complementary to the RNA transcript. Precipitated 5' end-labelled primer (10-50,000 c.p.m.) was resuspended in 5 µl of 120 mM NaCl containing the RNA sample to be mapped plus carrier yeast RNA to 20 µg. Samples were sealed into 20 µl capillaries, denatured for 5 mins at 85°C and allowed to hybridize for 1-2 hrs at 60°C. The capillary contents were then evacuated into a microfuge tube and 5 µl of primer extension cocktail (each dNTP at 3 mM, 100 mM Tris-Cl pH 8.3, 12 mM MaOAc, 100 U/ml reverse transcriptase, 20 mM DTT) added. These 10 µl mixes were incubated 1 hr at 45°C, the reaction stopped by addition of 5 µl formamide dye and the tubes were boiled 5-10 mins before loading the samples onto a 6% sequencing gel. The gel was run 1½-2 hrs and autoradiographed at -70°C with a screen for 15-72 hrs.

Primer extension was also used to sequence the RNA hybridized to the DNA primer (Donoghue and Hunter, 1982). Initial 20 µl hybrid-

izations using 4-10 times more primer and RNA were divided equally between 4 labelled microfuge tubes labelled A, G, C and T. To each tube was added 4 µl of an appropriate sequencing mix, detailed in Table 5, with 1 µl (1 U) reverse transcriptase. Incubation was for 1 hr at 45°C and the samples were boiled with formamide dye and run on a sequencing gel as before.

Table 5. Primer extension sequencing mixes.

MIX	µl each dNTP at 10 mM	µl 0.5 mM ddNTP	µl buffer*	µl DW	Final $\frac{[ddNTP]}{[dNTP]}$ ratio
A	1 x 4	26	18	32	1.3
G	1 x 4	6	18	52	0.3
C	1 x 4	8	18	50	0.4
T	1 x 4	18	18	40	0.9

*18 µl buffer comprises 10 µl 1 M Tris-Cl, pH 8.3 (50 mM), 4 µl 300 mM MgOAc (6 mM) and 4 µl 500 mM DTT (10 mM). Numbers in parentheses indicate the final concentrations of these reagents in the sequencing reaction.

b) S1 mapping (Favaloro et al., 1980).

S1 mapping can be used to map 5' and 3' ends of RNA transcripts using ds DNA probes with the strand complementary to the RNA labelled at its 5' or 3' end respectively.

Labelled DNA probe (20-50,000 c.p.m.) was precipitated with each RNA sample, plus yeast carrier RNA to 25 µg, for 1 hr at -70°C or o/n at -20°C. The precipitates were pelleted, washed and thoroughly dried before resuspending in 20 µl formamide buffer (40 mM PIPES pH 6.4, 400 mM NaCl, 1 mM EDTA, 80% deionized formamide). The microfuge tubes were then mounted in a polystyrene float and incubated for 15 mins at 85°C, followed by 52°C o/n. Samples were diluted with 300 µl S1 digestion buffer (280 mM NaCl, 30 mM NaOAc, pH 4.4, 4.5 mM $ZnSO_4$, 20 µg/ml ss DNA, 100-400 U/ml S1 nuclease) and maintained at 37°C for

1-2 hrs. Reactions were terminated by the addition of EDTA to 10 mM and the RNA/DNA hybrids precipitated on dry ice with the addition of NH_4OAc to 0.5 M, 20 µg carrier RNA and 550 µl isopropanol. The pelleted hybrids were redissolved in 5 µl formamide dye, boiled for 5 mins and loaded onto a 6% sequencing gel as for primer extension reactions.

c) Gel scanning.

Autoradiographs of primer extension and S1 mapping gels were scanned using a Joyce Loeble Chromoscan 3 densitometer to determine the relative intensities of the observed bands.

2.10 In Vitro Transcription (Jat et al., 1982).

The transcriptional capabilities of cloned genomic DNA fragments were assayed in HeLa cell extracts (Manley et al., 1980) kindly provided by Dr. G. Veldmann.

Particular subclones were digested with restriction enzyme(s), so separating the cloned insert from vector sequences, in order to prevent read-through from strong plasmid promoters during the transcription assay. This DNA was then used to 'programme' HeLa extracts that had been dialysed against 20 mM HEPES pH 7.9, 100 mM KCl, 12.5 mM $MgCl_2$, 0.1 mM EDTA, 2 mM DTT, 17% glycerol and stored in aliquots in liquid nitrogen. Standard incubations of 100 µl to 200 µl final volume contained each NTP at 0.5 mM, 10 mM creatine phosphate, 30-50 µg/ml DNA template and 50% HeLa extract, which was added last. Control incubations without template DNA and others containing the inhibitor α-amanitin were also set up. All samples were maintained at 30°C for 60-90 mins before terminating the reactions with 100 µl PK stop (100 mM Tris-Cl pH 7.5, 2% SDS, 20 mM EDTA, 400 µg/ml proteinase K) for 10 mins at 37°C. Samples were extracted once with 200 µl phenol mix and the nucleic acids were precipitated on dry ice by adjusting to 0.3 M NaOAc, pH 7.0 and adding 2½ volumes of ethanol. Pelleted nucleic acid was resuspended in 100 µl DNase buffer (10 mM $MgCl_2$, 50 mM Tris-Cl, pH 7.5) and RNase free DNase added to 20 µg/ml. Samples were incubated 30 mins at 37°C, then for a further 10 mins after addition of 100 µl PK stop and again phenol extracted. The RNA synthesized in the assays was recovered by precipitation and analysed by primer extension or S1 mapping.

REFERENCES

Ashburner, M., Chihara, C., Meltzer, P. and Richards, G. (1973) Cold Spring Harbor Symp. 38: 655-662.

Baldwin, T., Young, R. and Mindich, L. (1983) Focus 5, p. 14. Publ. Bethesda Research Laboratories, Inc.

Baltimore, D. (1981) Cell 24: 592-594.

Barta, A., Richards, R.I., Baxter, J.D. and Shine, J. (1981) Proc. Natl. Acad. Sci. U.S.A. 78: 4867-4871.

Benoist, C. and Chambon, P. (1981) Nature 290: 304-310.

Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucl. Acids Res. 8: 127-142.

Benton, W.D. and Dains, R.W. (1977) Science 196: 180-182.

Bernstein, L.B., Mount, S.M. and Weiner, A.M. (1983) Cell 32: 461-472.

Birnboim, H.C. and Doly, J. (1979) Nucl. Acids Res. 7: 1513-1523.

Bogenhagen, D.F. and Brown, D.D. (1981) Cell 24: 261-270.

Boss, M.A. (1983) Nature 303: 287-282.

Boyer, H.W. and Rouland-Dussoix, D. (1969) J. Mol. Biol. 41: 459-472.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. U.S.A. 75: 4853-4857.

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

Brinster, R.L., Chen, H.Y., Warren, R., Sarthy, A. and Palmiter, R.D. (1982) Nature 296: 39-42.

Brock, M.L. and Shapiro, D.J. (1983) Cell 34: 207-214.

Brutlag, D.L. (1980) Ann. Rev. Genet. 14: 121-144.

Brutlag, D.L., Clayton, J., Friedland, P. and Kedes, L.H. (1982) Nucl. Acids Res. 10: 279-294.

Buetti, E. and Diggelmann, H. (1983) EMBO J. 2: 1423-1429.

Busslinger, M., deBoer, E., Wright, S., Grosveld, F.G. and Flavell, R.A. (1983a) Nucl. Acids Res. 11: 3559-3569.

Busslinger, M., Hurst, J. and Flavell, R.A. (1983b) Cell 34: 197-206.

Calabretta, B., Robberson, D.L., Barrera-Saldana, H.A., Lambrou, T.P. and Saunders, G.F. (1982) Nature 296: 219-225.

Carter, D.B., Yamada, K. and Harris, S.E. (1983) Molecular and Cell. Endocrinol. 31: 199-214.

Cato, A.C.B. (1983) Biosci. Rep. 3: 101-111.

Chandler, V.L., Maler, B.A. and Yamamoto, K.R. (1983) Cell 33: 489-499.

Coffino, P. (1981) Nature 292: 492-493.

Compton, J.G., Schrader, W.T. and O'Malley, B. (1983) Proc. Natl. Acad. Sci. U.S.A. 80: 16-20.

Danner, D.B. and Pifer, M.L. (1982) Gene 18: 101-105.

Darnell, J.E. (1982) Nature 297: 365-371.

Davidson, E.H. and Britten, R.J. (1979) Science 204: 1052-1059.

Davidson, E.H. and Posakony, J.W. (1982) Nature 297: 633-635.

Davison, B.L., Mulvihill, E.R., Egly, J.M. and Chambon, P. (1982) Cold Spring Harbor Symp. 47: 965-975.

Dean, D.C., Knoll, B.J., Riser, M.E.and O'Malley, B.W. (1983) Nature 305: 551-554.

Denhardt, D.T. (1977) Biochem. Biophys. Res. Comm. 23: 641-646.

Denhardt, D.T., Dressler, D. and Ray, D.S. (1978) The single-stranded DNA phages. Cold Spring Harbor Laboratory.

Dierks, P., van Ooyen, A., Cochran, M.D., Dobkin, C., Reiser, J. and Weissmann, C. (1983) Cell 32: 695-706.

Doerfler, W. (1983) Ann. Rev. Biochem. 52: 93-124.

Donehower, L.A., Huang, A.L. and Hager, G.L. (1981) J. Virol. 37: 226-238.

Donoghue, D.J. and Hunter, T. (1982) Nucl. Acids Res. 10: 2549-2564.

Doolittle, W.F. and Sapienza, C. (1980) Nature 284: 601-603.

Dover, G. (1980) Nature 285: 618-620.

Duncan, C.H., Jagadeeswaran, P., Wang, R.C. and Weissman, S.M. (1981) Gene 13: 185-196.

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

Elder, J.T., Duncan, C.H. and Weissman, S.M. (1981) Nucl. Acids Res. 9: 1171-1189.

Everett, R.D., Baty, D. and Chambon, P. (1983) Nucl. Acids Res. 11: 2447-2464.

Fang, S. and Liao, S. (1971) J. Biol. Chem. 246: 16-24.

Favaloro, J., Treisman, R. and Kamen, R.I. (1980) Methods in Enzymology 65: 718-749.

Felsenfeld, G. and McGhee, J. (1982) Nature 296: 602-603.

Fitzgerald, M. and Shenk, T. (1981) Cell 24: 251-260.

Forsgren, B., Bjork, P., Carlstrom, K., Gustafsson, J-A., Pousette, A. and Hogberg, B. (1979) Proc. Natl. Acad. Sci. U.S.A. 76: 3149-3153.

Galli, G., Hofstetter, H. and Birnstiel, M.L. (1981) Nature 294: 626-631.

Garoff, H. and Ansorge, W. (1981) Analytical Biochem. 115: 450-457.

Geisse, S., Scheidereit, C., Westphal, H.M., Hynes, N.E., Groner, B. and Beato, M. (1982) EMBO J. 1: 1613-1619.

Georgiev, G.P., Kramerov, D.A., Ryskov, A.P., Skryabin, K.G. and Lukanidin, E.M. (1982) Cold Spring Harbor Symp. 47: 1109-1121.

Gorski, J. and Gannon, F. (1976) Ann. Rev. Physiol. 38: 425-450.

Govindan, M.V., Spiess, E. and Majors, J. (1982) Proc. Natl. Acad. Sci. U.S.A. 79: 5157-5161.

Gronemeyer, H. and Pongs, O. (1980) Proc. Natl. Acad. Sci. U.S.A. 77: 2108-2112.

Groner, B., Buetti, E., Diggelmann, H. and Hynes, N.E. (1980) J. Virol. 36: 734-745.

Grosveld, G.C., deBoer, E., Shewmaker, C.K. and Flavell, R.A. (1982) Nature 295: 120-126.

Grosveld, G.C., Shewmaker, C.K., Jat, P. and Flavell, R.A. (1981) Cell 25: 215-226.

Grunstein, M. and Hogness, D.S. (1975) Proc. Natl. Acad. Sci.U.S.A. 72: 3961-3965.

Hagenbüchle, O., Bovey, R. and Young, R.A. (1980) Cell 21: 179-187.

Hager, L.J. and Palmiter, R.D. (1981) Nature 291: 340-342.

Hamilton, T.H., Widnell, C.C. and Tata, J.R. (1965) Biochim. Biophys. Acta 108: 168-172.

Hattman, S., Brooks, J.E. and Masurekar, M. (1978) J. Mol. Biol. 126: 367-380.

Haynes, S.R. and Jelinek, W.R. (1981) Proc. Natl. Acad. Sci. U.S.A. 78: 6103-6134.

Hershko, A., Mamont, P., Shields, R. and Tomkins, G.M. (1971) Nature New Biol. 232: 206-211.

Heyns, W. and De Moor, P. (1977) Eur. J. Biochem. 78: 221-230.

Heyns, W., Peeters, B., Mous, J., Rombauts, W. and De Moor, P. (1978a) Eur. J. Biochem. 89: 181-186.

Heyns, W., Van Damme, B. and De Moor, P. (1978b) Endocrinology 103: 1090-1095.

Higgins, S.J. and Gehring, U. (1978) Adv. Cancer Res. 28: 313-397.

Higgins, S.J., Smith, S.E. and Wilson, J. (1982) Mol. Cell. Endocrinol. 27: 55-65.

Holmes, D.S. and Quigley, M. (1981) Anal. Biochem. 114: 193-197.

Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1979) J. Mol. Biol. 132: 289-306.

Huang, A.L., Ostrowski, M.C., Berard, D. and Hager, G.L. (1981) Cell 27: 245-255.

Hynes, N., van Ooyen, A.J.J., Kennedy, N., Herrlich, P., Ponta, H. and Groner, B. (1983) Proc. Natl. Acad. Sci. U.S.A. 80: 3637-3641.

Ish-Horowicz, D. and Burke, J.F. (1981) Nucl. Acids Res. 9: 2989-2998.

Jagadeeswaran, P., Forget, B.G. and Weissman, S.M. (1981) Cell 26: 141-142.

Jat, P., Roberts, J.W., Cowie, A. and Kamen, R. (1982) Nucl. Acids Res. 10: 871-886.

Jelinek, W.R. and Haynes, S.R. (1982) Cold Spring Harbor Symp. 47: 1123-1130.

Jelinek, W.R. and Schmid, C.W. (1982) Ann. Rev. Biochem. 51: 813-844.

Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Biro, P.A., Choudary, P.V., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L. and Schmid, C.W. (1980) Proc. Natl. Acad. Sci. U.S.A. 77: 1398-1402.

Jensen, E.V., Hurst, D.J., De Sombre, E.R. and Jungblut, P.W. (1967) Science 158: 385-387.

Jensen, E.V. and Jacobsen, H.I. (1962) Recent Progr. Hormone Res. 18: 387-414.

Jensen, E.V., Suzuki, T., Kawashima, T., Stumpf, W.E., Jungblut, P.W. and De Sombre, E.R. (1968) Proc. Natl. Acad. Sci. U.S.A. 59: 632-638.

Jung, A., Sipple, A.E., Grez, M. and Schütz, G. (1980) Proc. Natl. Acad. Sci. U.S.A. 77: 5759-5763.

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

King, R.J.B., Cambray, G.J. and Robinson, J.H. (1976) J. Steroid Biochem. 7: 869-873.

Kunnath, L. and Locker, J. (1983) EMBO J. 2: 317-324.

Kurtz, D.T., McCullough, L., Bishop, D.K. and Manos, M.M. (1982) Cold Spring Harbor Symp. 47: 985-988.

Laskey, R. and Mills, A.D. (1977) FEBS Letts. 82: 314-316.

Lea, O.A., Petrusz, P. and French, F.S. (1979) J. Biol. Chem. 254: 6196-6202.

Lee, F., Mulligan, R., Berg, P. and Ringold, G. (1981) Nature 294: 228-232.

Leinwand, L.A., Wydro, R.M. and Nadal-Ginard, B. (1982) Mol. Cell. Biol. 2: 1320-1330.

Lemischka, I. and Sharp, P.A. (1982) Nature 300: 330-335.

Lesser, B. and Bruchovsky, N. (1973) Biochim. Biophys. Acta 308: 426-437.

Maat, J., van Boveren, C.P. and van Ormondt, H. (1980) Gene 10: 27-38.

Maizels, N. (1976) Cell 9: 431-438.

Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J., O'Connell, C., Quorn, D., Sim, G.K. and Efstratiadis, A. (1978) Cell 15: 687-701.

Manley, J.L., Fire, A., Cano, A., Sharp, P.A. and Gefter, M.L. (1980) Proc. Natl. Acad. Sci. U.S.A. 77: 3855-3859.

Mathis, D.J. and Chambon, P. (1981) Nature 290: 310-315.

Maxam, A. and Gilbert, W. (1980) Methods in Enzymol. 65: 499-560.

Maxwell, I.H., Maxwell, F. and Hahn, W.E. (1977) Nucl. Acids Res. 4: 241-246.

Mayo, K.E. and Palmiter, R.D. (1981) J. Biol. Chem. 256: 2621-2624.

McDonell, M.W., Simon, M.N. and Studier, F.W. (1977) J. Mol. Biol. 110: 119-146.

McKnight, G.S. and Palmiter, R.D.(1979) J. Biol. Chem. 254: 9050-9058.

McKnight, S.L. (1982) Cell 31: 355-365.

McKnight, S.L., Gavis, E.R., Kingsbury, R. and Axel, R. (1981) Cell 25: 385-398.

Means, A.R. and Hamilton, T.H. (1966) Proc. Natl. Acad. Sci.U.S.A. 56: 686-693.

Messing, J. (1979) Recomb. DNA Tech. Bull. 2: 43-48.

Messing, J., Crea, R. and Seeburg, P.H. (1981) Nucl. Acids Res. 9: 309-321.

Messing, J. and Vierira, J. (1982) Gene 19: 269-276.

Mirault, M-E., Southgate, R. and Delwart, E. (1982) EMBO J. 1: 1279-1285.

Mount, S.M. (1982) Nucl. Acids Res. 10: 459-472.

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

Mulvihill, E.R., Le Pennec, J-P. and Chambon, P. (1982) Cell 24: 621-632.

Murray, N.E., Brammar, W.J. and Murray, K. (1977) Mol. Gen. Genet. 150: 53-61.

Nevins, J.R. (1983) Ann. Rev. Biochem. 52: 441-466.

Nordheim, A. and Rich, A. (1983) Nature 303: 647-678.

Orgel, L.E. and Crick, F.H.C. (1980) Nature 284: 604-607.

Orkin, S.H., Sexton, J.P., Cheng, T., Goff, S.C., Giardina, P.J.V., Lee, J.I. and Kazazian, H.H. (1983) Nucl. Acids Res. 11: 4727-4734.

Page, G.S., Smith, S. and Goodman, H.M. (1981) Nucl. Acids Res. 9: 2084-2104.

Page, M.J. and Parker, M.G. (1982) Mol. Cell. Endocrinol. 27: 343-355.

Page, M. and Parker, M. (1983) Cell 32: 495-502.

Palmiter, R.D. and Carey, N.H. (1974) Proc. Natl. Acad. Sci. U.S.A. 71: 2357-2361.

Paolella, G., Lucero, M.A., Murphy, M.H. and Baralle, F.E. (1983) EMBO J. 2: 691-696.

Parker, M.G. and Mainwaring, W.I.P. (1977) Cell 12: 401-407.

Parker, M.G. and Mainwaring, W.I.P. (1978) J. Steroid Biochem. 9: 455-460.

Parker, M., Needham, M. and White, R. (1982a) Nature 298: 92-94.

Parker, M., Needham, M., White, R., Hurst, H. and Page, M. (1982b) Nucl. Acids Res. 10: 5121-5132.

Parker, M.G. and Scrace, G.T. (1978) Eur. J. Biochem. 85: 399-406.

Parker, M.G. and Scrace, G.T. (1979) Proc. Natl. Acad. Sci.U.S.A. 76: 1580-1584.

Parker, M.G., Scrace, G.T. and Mainwaring, W.I.P. (1978) Biochem. J. 170: 115-121.

Parker, M.G., White, R., Hurst, H., Needham, M. and Tilley, R. (1983) J. Biol. Chem. 258: 12-15.

Parker, M.G., White, R. and Williams, J.G. (1980) J. Biol. Chem. 255: 6996-7001.

Payvar, F., Firestone, G.L., Ross, S.R., Chandler, V.L., Wrange, O., Carlstedt-Duke, J., Gustafsson, J-A. and Yamamoto, K.R. (1982) J. Cell. Biochem. 19: 241-247.

Pearson, W.R., Wu, J-R. and Bonner, J. (1978) Biochemistry 17: 51-59.

Peeters, B., Rombauts, W., Mous, J. and Heyns, W. (1981) J. Biochem. 115: 115-121.

Pelham, H.R.B. (1982) Cell 30: 517-528.

Pfahl, M. (1982) Cell 31: 475-482.

Price, D. and Williams-Ashman, H.G. (1961) In: 'Sex and Internal Secretion' (ed. W.C. Young) 3rd Edition, pp. 366-448. Williams and Wilkins, Baltimore, Maryland, U.S.A.

Proudfoot, N. and Brownlee, G.G. (1976) Nature 263: 211-214.

Proudfoot, N.J., Shander, M.H.M., Manley, J.L., Geflet, M.L. and Maniatis, T. (1980) Science 209: 1329-1336.

Razin, A. and Riggs, A.D. (1980) Science 210: 604-610.

Renkawitz, R., Beug, H., Graf, T., Matthias, P., Grez, M. and Schütz, G. (1982) Cell 31: 167-176.

Renkawitz, R., Danesch, U., Matthias, P. and Schütz, G. (1983) J. Steroid Biochem. (in press).

Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113: 237-251.

Ringold, G.M., Yamamoto, K.R., Bishop, J.M. and Varmus, H.E. (1977) Proc. Natl. Acad. Sci. U.S.A. 74: 2879-2883.

Robins, D.M., Paek, I., Seeburg, P.H. and Axel, R. (1982) Cell 29: 623-631.

Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13: 319-353.

Sanger, F. and Coulson, A.R. (1978) FEBS Letts. 87: 107-110.

Sargent, T.D., Wu, J-R., Sala-Trepat, J.M., Wallace, R.B., Reyes, A.A. and Bonner, J. (1979) Proc. Natl. Acad. Sci. U.S.A. 76: 3256-3260.

Scheidereit, C., Geisse, S., Westphal, H.M. and Beato, M. (1983) Nature 304: 749-752.

Schrader, W.T., Birnbaumer, M.E., Hughes, M.R., Weigel, N.L., Grody, W.W. and O'Malley, B.W. (1981) Recent Progr. Hormone Res. 37: 583-633.

Schrader, W.T., Compton, J.G. and O'Malley, B.W. (1983) J. Steroid Biochem. (in press).

Sharp, P.A. (1981) Cell 23: 643-646.

Sharp, P.A. (1983) Nature 301: 471-472.

Shen, C-H.J. and Maniatis, T. (1982) J. Mol. Appl. Genet. 1: 343-360.

Siebenlist, V., Simpson, R.B. and Gilbert, W. (1980) Cell 20: 269-281.

Southern, E.M. (1975) J. Mol. Biol. 98: 503-517.

Southern, E.M. (1980) Methods in Enzymol. 68: 152-175.

Spritz, R.A. and Forget, B.G. (1983) Am. J. Hum. Genet. 35: 333-361.

Staden, R. (1980) Nucl. Acids Res. 8: 3673-3694.

Sutcliffe, J.G., Milner, R.J., Bloom, F.E. and Lerner, R.A. (1982) Proc. Natl. Acad. Sci. U.S.A. 79: 4942-4946.

Swaneck, G.E., Nordstrom, J.L., Kreuzaler, F., Tsai, M-J. and O'Malley, B.W. (1979) Proc. Natl. Acad. Sci. U.S.A. 76: 1049-1053.

Tavernier, J., Gheysen, D., Duerinck, F., Van der Heyden, J. and Fiers, W. (1983) Nature 301: 634-636.

Tjian, R. (1981) Cell 26: 1-2.

Toft, D. and Gorski, J. (1966) Proc. Natl. Acad. Sci. U.S.A. 55: 1574-1581.

Twigg, A.J. and Sherratt, D. (1980) Nature 283: 216-218.

Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T. and Gesteland, R.F. (1981) Cell 26: 11-17.

Vardimon, L., Kressmann, A., Cedar, H., Maechler, M. and Doerfler, W. (1982) Proc. Natl. Acad. Sci. U.S.A. 79: 1073-1077.

Wahli, W., Wyler, T., Weber, R. and Ryfell, G.U. (1978) Eur. J. Biochem. 86: 225-234.

Walter, P. and Blobel, G. (1982) Nature 299: 691-698.

Wang, A.H-J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Broom, J.H., van der Marel, G. and Rich, A. (1979) Nature 282: 680-686.

Wasylyk, B., Derbyshire, R., Guy, A., Molko, D., Roget, A., Teoule, R. and Chambon, P. (1980) Proc. Natl. Acad. Sci. U.S.A. 77: 7024-7028.

Weaver, R.F. and Wiessmann, C. (1979) Nucl. Acids Res. 7: 1175-1193.

Weinmann, R., Raskas, H.J. and Roeder, R.G. (1974) Proc. Natl. Acad. Sci. U.S.A. 71: 3426-3430.

White, R. and Parker, M. (1983) J. Biol. Chem. 258: 8943-8948.

Wilkes, M.M., Pearson, W.R., Wu, J-R. and Bonner, J. (1978) Biochemistry 17: 61-69.

Wilson, J.D. and Gloyna, R.E. (1970) Recent Progr. Hormone Res. 26: 309-336.

Witney, F.R., Lieberman, R. and Furano, A.V. (1983) Fed. Proc. 42: 2264.

Young, H.A., Shih, T.Y., Scolnick, E.M. and Parks, W.P. (1977) J. Virol. 21: 139-146.

Zarucki-Schulz, T., Tsai, S.Y., Itakura, K., Soberon, X., Wallace, R.B., Tsai, M-J., Woo, S.L.C. and O'Malley, B.W. (1982) J. Biol. Chem. 257: 11070-11077.

Zieve, G.W. (1981) Cell 25: 296-297.

Additional reference:

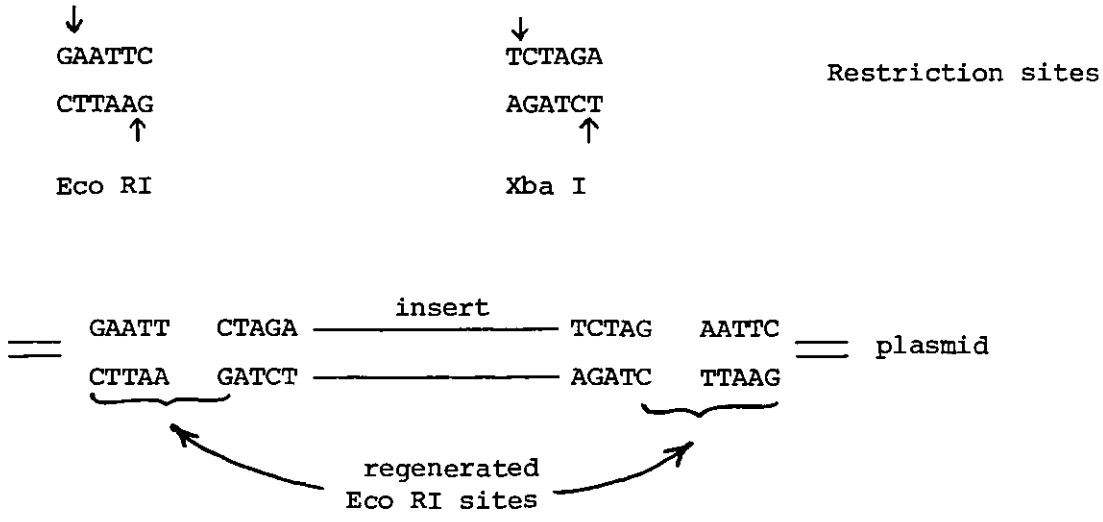
Feinberg, R.F., Sun, L-H.K., Ordahl, C.P. and Frankel, F.R. (1983) Proc. Natl. Acad. Sci. U.S.A. 80: 5042-5046.

APPENDIX: DETAILS OF GENOMIC SUBCLONES

1. Subclones in pAT 153

All the genomic subclones of C3(1) and C3(2) used in this thesis are shown diagrammatically in Fig. 28. Cloned inserts obtained in both orientations were designated A (5'→3' clockwise) and B (5'→3' anti-clockwise). Most of these subclones were straightforward to produce: p61 and p81 were cloned into the Eco RI site of pAT 153 and selected using tetracycline resistance; p5A1 and p63 were cloned into the Bam HI site and selected using ampicillin; p11B1, 2 and 3 were cloned into the Pst I site and selected using tetracycline.

Both p62 and p5A2 were subcloned into the Eco RI site of pAT 153. In each case the 2.6 kbp Xba I fragment was obtained by digestion of larger subclones blunt-ended and cloned into blunt-ended Eco RI digested vector. This approach was adopted in order to regenerate an Eco RI site each side of the insert, as shown in the diagram:-



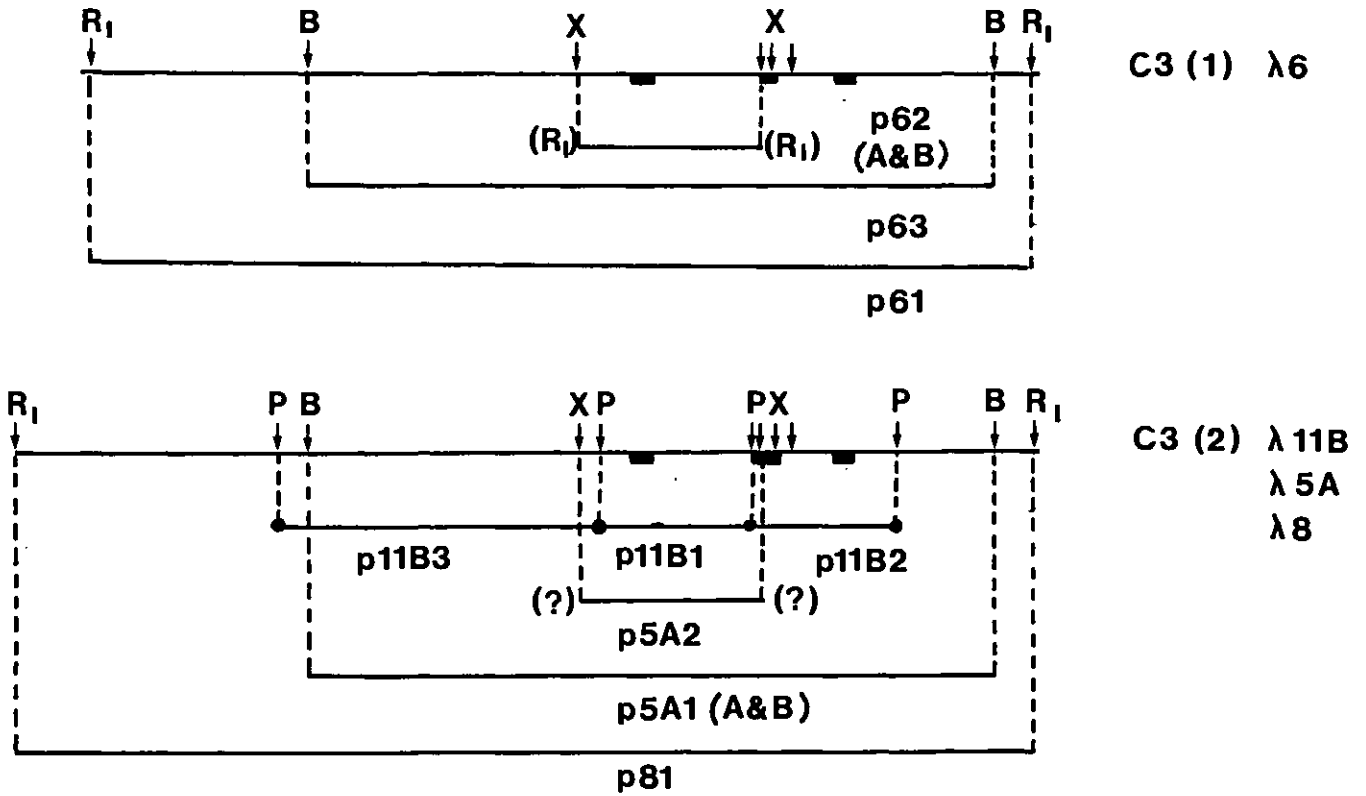
Cut and filled-in restriction sites in a blunt-ended ligation of plasmid to insert.

This only worked for p62 when Klenow DNA polymerase was used in the blunt-ending steps. for p5A2 intact DNA polymerase I was used which retains the 5' to 3' exonuclease activity, hence can nibble away at the DNA 5' overhangs generated by the enzyme digestions, rather than

Figure 28. Subclones of the C3 genes.

The diagrams show simple restriction enzyme maps of the sequences isolated as genomic λ-clones.
B = Bam HI, P = Pst I, RI = Eco RI, and X = Xba I.
Exons are shown as blocks. Those portions of the genomic clones that were individually subcloned into pAT 153 are indicated beneath each restriction map, together with a name derived from the parent λ-clone.

The clone p5A2 has (?) at each end because restriction enzyme sites were not regenerated at the ends of the insert during cloning (see text).


R1
B
X
X
B
R1
(R1)
(R1)
p62
(A&B)
p63
p61
C3 (1) λ6
R1
P
B
X
P
PX
P
B
R1
p11B3
p11B1
p11B2
(?)
(?)
p5A2
p5A1 (A&B)
p81
C3 (2) λ 11B
λ 5A
λ 8

fill them in. For this reason the insert in p5A2 is not flanked by restriction enzyme sites and cannot be cut out from the vector. Luckily the insert was in the correct orientation (A) to be used in Maxam and Gilbert sequencing of the beginning of the second exon (see R2).

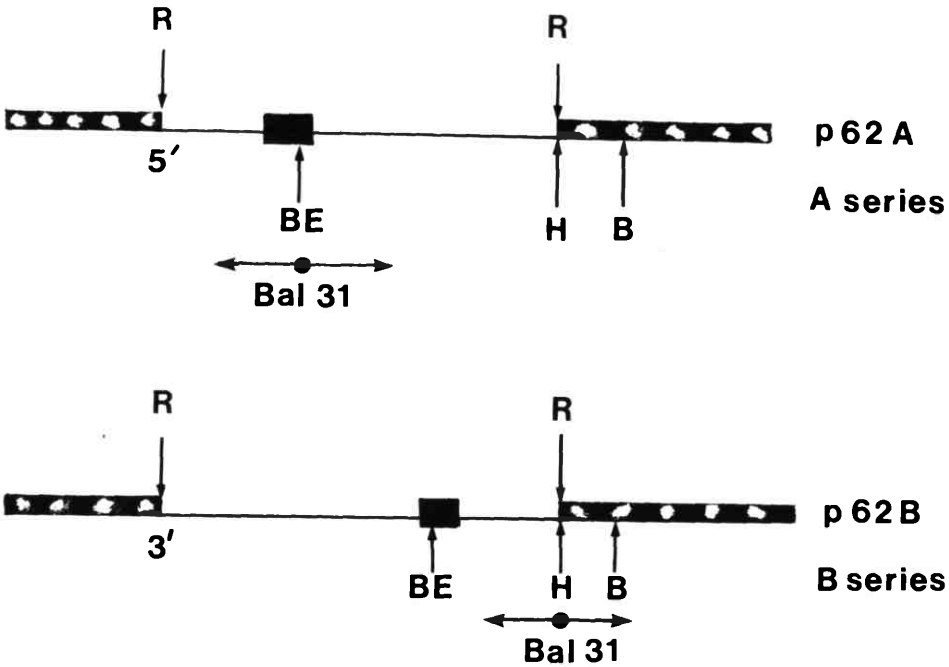
It should be noted that the middle Xba I site of the cluster of three around the second exon of both genes cannot be digested when plasmids were grown in HB101 or LE392 cells and the site was only found by sequencing. This revealed that the Xba I site is within the sequence TCTAGATC. As the strains of bacteria used contain the dam methylase which introduces methyl groups at position N^6 of all adenine residues present in sequences 5' GATC 3' (Hattman et al., 1978) a methyl group will be present on the final A of this Xba I recognition sequence, which apparently inhibits the enzyme.

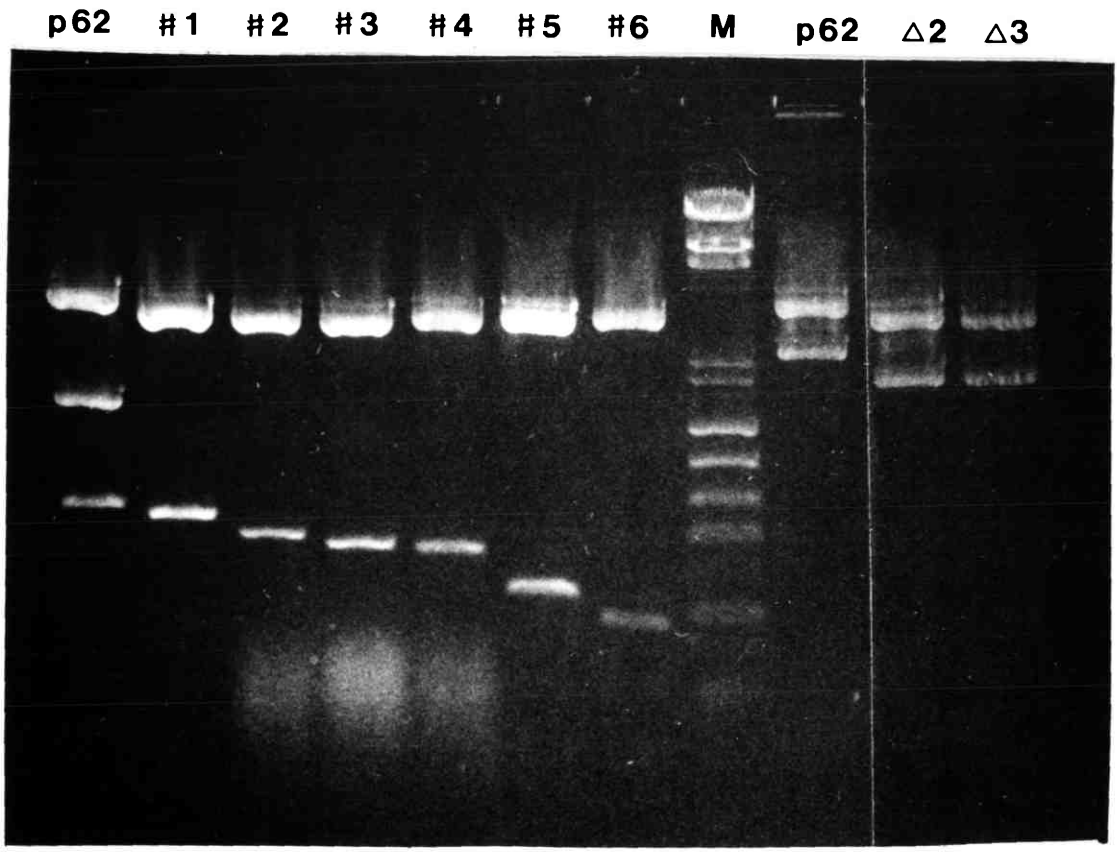
2. 5' Deletion Mutants of C3(1)

Deletion mutants around the first exon and 5' flanking DNA of C3(1) were generated using Bal 31 nuclease on the subclones p62A and B, as shown diagrammatically in Fig. 29. The A series of deletions remaining proximal exon and flanking sequences were made using p62A by resecting from the Bst E_{II} site within the first exon. The B series of deletions, remaining distal flanking DNA were made using p62B by resecting from the Hind III site in the plasmid. Suitably deleted molecules were then digested with Bam HI so cutting away unwanted regions of the insert, blunt-ended using Klenow DNA polymerase and recircularized with Bam linkers. The DNA was transfected into competent HB101 cells and selected for with ampicillin. Colonies from these first plates were then replica-plated onto ampicillin and tetracycline plates and the DNA from several Amp^R Tet^S colonies examined in plasmid mini preps for the size of the residual inserts. Suitable deletion mutants were cloned directly into Eco RI/Bam HI digested M13 mp 8 and 9 for sequencing, but were also grown up as large scale plasmid preparations for sizing (Fig. 29). Those that could be sequenced (see R2) allowed the precise deletion endpoints relative to C3 transcription at +1 to be determined:-

Figure 29. Generation of C3(1) 5' deletion mutants.

The two diagrams illustrate the strategy used to generate two series (A and B) of overlapping deletion mutants around the first exon and 5' flanking DNA of C3(1). The photograph shows digests of the parent subclone, p62, together with the deletion mutants obtained run on a 0.8% agarose gel. On the left hand side, p62 was digested with Bst E_{II} and Eco RI while the clones #1-6 were digested with Eco RI and Bam HI, the deletion end point. Similarly, on the right hand side, p62 was digested with Eco RI while the two clones Δ2 and Δ3 were treated with Eco RI and Bam HI. The markers were Hind III digested λ together with Hpa II digested polyoma DNA.


R
R
5′
BE
H
B
p62A
A series
Bal 31
R
R
3′
BE
H
B
p62B
B series
Bal 31


p62
#1
#2
#3
#4
#5
#6
M
p62
△2
△3

A series (3' deletions)		B series (5' deletions)	
#1	+32 A	Δ1	?
#2	-42 A	Δ2	?
#3	-78 A	Δ3	-195 B
#4	-116 A		
#5	-292 A		
#6	-383 A		

3. Subcloning into M13

The overall strategy for sequencing the two C3 genes is shown in Fig. 8. Below is a list of most of the clones derived in M13 vectors for this purpose, indicating the original plasmid subclone used as a DNA source. In general specific restriction fragments were isolated from small acrylamide gels and cloned singly into suitably restricted M13 vectors.

cs = coding strand clone
ncs = non coding strand clone.

Clones used to sequence C3(2):-

p11B1 2.1 kbp Pst I insert into mp 7 (3 clones only, 1cs and 2ncs).

p11B2 600 bp and 1000 bp Pst I/Hind III inserts into mp 8 and 9 (many ncs and cs clones).

p11B3 4.5 kbp Pst I insert isolated, digested with Sau 3A and shotgun cloned into Pst I/Bam HI cut mp 8 and 9. (Only cs clones in mp 8 were isolated, see R2).

All the other clones were derived from the 'shotgun' library (Hae III, Alu I and Sau 3A digests of p11B1 and p11B2 in mp 7).

Clones used to sequence C3(1):-

p61 900 bp and 700 bp Pst I/Bgl II fragments corresponding to the second and third exons into mp 8 and 9. 350 bp Xba I fragment, corresponding to middle exon, blunt-ended and cloned into blunt-ended Eco RI site of mp 9. Clones in both orientations were obtained.

p63 Dde I cut, blunt-ended then Bgl II cut. Fragments in region of 150 bp isolated and shotgun cloned into Hind II/Bam HI cut mp 8. 3 cs clones isolated.

Alu I cut, Bgl II cut fragments of ~300 bp cloned into Hind II/Bam H1 cut mp 8. 4 ncs clones isolated.

The above cloning ensured that sequences from the second and third exons were obtained from both strands (except for the few bases at the beginning of exon 2 that were determined using Maxam and Gilbert sequencing, see R2). To obtain the first exon and proximal 5' flanking sequences on both strands:-

p62 Isolated 2.6 kbp Eco RI insert, digested with Sau 3A and shotgun cloned into Eco RI/Bam HI cut mp 8 and 9. Obtained only large clones (~1.0 kbp) in mp 8 of ncs, so could sequence from first intron into the exon.

A series Further cs sequences determined using Eco RI/Bam HI inserts of the deletion mutants which would only clone into mp 8.

B series The smallest B deletion mutant ($\Delta 3$) cloned into Eco RI/Bam HI cut mp 8 to give the only cs clones and consequently non coding strand sequence obtained in this region.