Organization and Expression of the Mouse Surfeit Gene Cluster

Clare Marjory Pease Huxley

Thesis presented in partial fulfilment
of the degree of Doctor of Philosophy
to the University of London

Imperial Cancer Research Fund
Laboratories
Lincoln's Inn Fields
London, WC2A 3PX

Department of Biochemistry
Imperial College of Science
and Technology
London, SW7 2AZ

May 1989

Abstract

The Mouse Surfeit gene cluster has previously been identified as a very tight cluster of at least 4 housekeeping genes (Surf-1 to -4) with an organization which is unique among known mammalian genes, as no more than 73bp separates adjacent genes (Williams et al 1988). These four genes are organized as closely apposed alternating transcription units covering about 24kb of genomic DNA.

The Surf-1 and Surf-2 genes have been characterized by DNA sequence and transcription analysis (Williams 1986). In this work the Surf-3 and Surf-4 genes have been characterized, other Surfeit genes have been identified and CpG-rich islands have been located.

The Surf-4 gene has been cloned and sequenced. This gene is arranged as 6 exons spanning 14kb of DNA and encodes a putative protein of 269 amino acids. The two mRNAs specified by Surf-4 differ in the position of polyadenylation. The upstream region of the Surf-3 gene has been sequenced and the start of transcription determined by primer extension. Surf-3 is a member of a multigene family whose other members are shown to be predominantly processed pseudogenes. Two new genes have been identified in the Surfeit cluster bringing the total number of Surf genes to six.

Four CpG-rich islands clustered in 32kb of genomic DNA have been identified. The two flanking CpG-rich islands have been located 500kb and 100kb on either side of the cluster of 4 islands. This indicates that we have identified the full extent of clustered genes and islands.

The steady state level of the Surf-2, -3 and -4 mRNAs are not cell cycle regulated indicating that this type of temporal control is not responsible for the arrangement of the genes.

The Surfeit locus has been mapped close to the centromere of mouse chromosome 2, it lies approximately 45cM proximal to *β2-microglobulin* and about 6cM from the proto-oncogene *c-abl*.

Acknowledgements

Thanks to Mike for this thesis.

Thanks to Jeremy for many delicious meals.

Thanks to the everyone in Friedland for three very enjoyable years.

Contents

		Page no.
Title Page		1
Abstract		2
Acknowledgements		3
Contents		4
Figures and Tables		7
Abbreviations		9
Chapter 1:	**Introduction**	
1.1	Sharing of Regulatory Regions and Interactions Between Adjacent Genes	10
	Do these mechanisms function in eukaryotic cells	13
1.2	Organization of the Mammalian Genome	16
	Cistron like arrangements	16
	Genes are individually regulated	17
	Highly repetitive dispersed DNA	19
	CpG-rich islands	22
	Using the presence of CpG-rich islands to map genes	26
	Bidirectional promoters in CpG-rich islands	28
	Clusters of repeated genes	31
	Clusters of unrelated genes in mammalian genomes	35
	Clustered genes in drosophila	37
	Domains containing coregulated genes	39
	Distribution of coregulated genes	41
	Conclusion	42
1.3	Cloning the Mouse Surfeit Gene Cluster	44
1.4	Organization of the Mouse Surfeit Gene Cluster	47
1.5	Significance of the Clustering of the Surfeit Genes	49
Chapter 2:	**Structural Analysis of Surf-3**	
2.1	Introduction	50
2.2	Results: Sequence and Organization of Surf-3	51
2.3	Results: Cloning and Sequencing Surf-3 Pseudogenes	62
2.4	Discussion: What is the Surf-3 Protein?	67
2.5	Discussion: The TRK-h2 Oncogene	72
2.6	Further Investigations on Surf-3	73

Chapter 3: Structural Analysis of Surf-4

3.1 Introduction 77
3.2 Results: Genomic and cDNA Clones of Surf-4 78
3.3 Results: Locating the 5' End of Surf-4 89
3.4 Results: The Two Surf-4 Messages Differ in their Poly(A) Addition Site 95
3.5 Conclusions and Discussion 98

Chapter 4: CpG-Rich Islands

4.1 Introduction 100
4.2 Results: Four CpG-rich Islands in the Surfeit Cluster 101
4.3 Results: Location of the Neighbouring CpG-rich Islands 112
4.4 Results: Location of CpG-rich Islands Neighbouring Ribosomal Protein Genes 115
4.5 Discussion 122

Chapter 5: Identification of Two New Genes and Repetitive Elements in the Surfeit Cluster

5.1 Introduction 123
5.2 Results: Two New Genes 123
5.3 Results: Repetitive Elements 131
5.4 Discussion and Further Work 140

Chapter 6: Cell Cycle Regulation of the Surfeit Genes

6.1 Introduction 141
6.2 Results 143
6.3 Discussion 145

Chapter 7: Mapping

7.1 Introduction 154
7.2 Results: Mapping the Surfeit Cluster and *c-abl* 156
7.3 Results: The Surfeit Cluster and *c-abl* cannot be Linked with Pulsed Field gel Electrophoresis 164
7.4 Discussion 168

Chapter.8: Conclusion and discussion

8.1 Summary of Results 170
8.2 The Surfeit Cluster is a Novel Arrangement of Genes 172
8.3 How to resolve Chance from Function 176
8.4 Conclusion 179

Chapter 9: Materials and Methods

9.1 Materials 180

9.2 Methods 185

Appendix 1a Map and Sequencing of Surf-3 206

Appendix 1b Maps and Sequencing of Surf-3 cDNAs 207

Appendix 1c Maps and Sequencing of Surf-3 pseudogenes 208

Appendix 2a Map of Surf-4 209

Appendix 2b Maps and Sequencing of Surf-4 cDNAs 210

Appendix 2c Maps and Sequencing of genomic exons of Surf-4 211

Appendix 3 Map of the Surf-5 and -6 region 212

Appendix 4a Protein Consensus Sequences 213

Appendix 4b Transcription factor binding sites 215

References. 218

Figures and Tables

Figure		Page no.
1.1	Isolation of the MES-1 Expression Sequence	46
1.2	Organization of the Mouse Surfeit Gene Cluster	48
2.1	Northern and Southern blot analysis of Surf-3	54
2.2	DNA sequence and protein encoded by Surf-3	56
2.3	Primer extension analysis of the 5' end of Surf-3	58
2.4	RNase protection analysis of the 5' end of Surf-3	60
2.5	The 5' ends of three pseudogenes of Surf-3	64
2.6	Three completely sequenced pseudogenes of Surf-3	65
2.7	Comparison of the DNA sequences of the 5' ends of mammalian ribosomal protein genes	70
3.1	Organization of Surf-4 and the Surfeit cluster	82
3.2	Southern and Northern blots of the Surf-4	84
3.3	Sequence of Surf-4	86
3.4	Primer extension and S1 mapping to determine the 5' end of surf-4	91
3.5	DNA sequence of the upstream region of Surf-4	93
3.6	S1 mapping to determine the nature of the difference between the 2.0 and 2.8 kb Surf-4 mRNAs	96
4.1	CpG dinucleotide content of the Surfeit locus	101
4.2	Location of four CpG-rich islands in the Surfeit cluster	106
4.3	Further analysis of the CpG-rich islands at the far left of the Surfeit cluster and at the 5' end of Surf-4	108
4.4	Southern blot of methylated BssHII cleavage site	110
4.5	CpG-rich islands neighbouring the Surfeit Cluster	113
4.6	Long range map around the ribosomal protein L30 gene	118
4.7	Long range map around the ribosomal protein L32 gene	120

5.1	Map of the 8 kb of DNA upstream of Surf-3	126
5.2	The Surf-5 transcription unit	127
5.3	The Surf-6 transcription unit	129
5.4	Low level repetitive DNA upstream of Surf-3	132
5.5	B1 and B2 elements in the Surfeit Cluster	134
5.6	B1 element in the Surfeit cluster	137
5.7	Two B2 elements in the Surfeit cluster	138
5.8	Adjacent B2 elements upstream of Surf-4	139
6.1	Northern blots of cell cycle dependent expression of the Surfeit genes	146
6.2	Graphs showing the cell cycle sorting of the HeLa cells	148
6.3	Expression of Surf-2 and Surf-3 through the cell cycle	150
6.4	Expression of Surf-4 through the cell cycle	152
7.1	Map of mouse chromosome 2	162
7.2	Pulsed field gel hybridized with Surfeit probes and *v-abl*	166

Table Page no.

1.1	Island specific rare cutting enzymes	26
4.1	C and G content of the four Surfeit CpG-rich islands	104
7.1	Distribution of DNA polymorphisms in an interspecific backcross of 34 animals	159
7.2	Distribution of DNA polymorphisms in an interspecific backcross of 54 animals	160
7.3	Genetic distances between markers	161

Abbreviations

A	adenine
bp	base pair(s)
BSA	bovine serum albumin
C	cytosine
cM	centiMorgan
cDNA	complementary DNA
cpm	counts per minute
DMEM	Dulbecco modified Eagles medium
DNA	deoxyribonucleic acid
DTT	dithiothreitol
EDTA	diaminoethane tetraacetic acid (disodium salt)
FCS	foetal calf serum
G	guanine
HEPES	N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid
hrs	hours
IPTG	isopropyl-1-thio-β-D-galactoside
kb	kilobase(s)
kDa	kiloDalton(s)
LMA	low melting point agarose
Mb	mega base pairs
mins	minutes
MOPS	3-[N-Morpholino]propanesulphonic acid
mRNA	messenger RNA
nt	nucleotide(s)
o/n	overnight
ORF	open reading frame
pfu	plaque forming units
PIPES	Piperazine-N,N'-bis[2-ethanesulphonic acid]
RNA	ribonucleic acid
RT	room temperature
SDS	sodium dodecyl sulphate
secs	seconds
Tris	Tris[hydroxymethyl]aminomethane

Chapter 1: Introduction

1.1 Sharing of Regulatory Regions and Interactions Between Adjacent Genes

An organism requires the ability to coregulate classes of genes and to regulate genes in a fashion which is dependent on the expression of other genes. This can be achieved by transacting transcription factors which recognize and bind specific DNA sequences near, or within, the gene which is to be regulated. It can also be achieved very efficiently, in terms of the amount of DNA specifying regulatory functions, if genes share regulatory elements or if the expression of one gene directly inhibits the transcription or translation of another gene.

There are a number of molecular mechanisms whereby the *cis* location of genes can be efficiently utilized towards the regulation of their expression. The location of numerous genes on a single cistronic mRNA allows coregulation of all the genes in the cistron. Bidirectional promoters driving transcription of two divergent genes allows the possible sharing of regulatory elements. Shared regulatory elements can be used either to co-ordinately regulate, or to give alternate expression of the two genes. Tandem location of two promoters can cause promoter occlusion such that when the upstream promoter is active the downstream one is inactivated. Convergent transcription of two genes allows interference of the RNA polymerase of one gene with the transcriptional machinery of the other gene. Convergent transcription also allows antisense control if the two RNAs hybridize to give double stranded RNA which is not functional. Finally, regulatory elements can control the chromatin conformation of domains of DNA thus coregulating all the genes in that region.

Prokaryotes such as *E. coli* have been shown to utilize all the above methods to achieve gene control, except the use of domains of DNA in

different chromatin conformations. The cistronic arrangement of genes is very common in prokaryotes, but the other mechanisms vary in the degree to which they are used.

Over twenty bidirectional promoters have been found in the chromosome of *E.coli* and more than 40 have been found in prokaryotic viruses, transposable elements, mitochondria and chloroplasts (Beck & Warren 1988). The *bioA - bioBFCD* system is an example of structural genes involved in the same biochemical pathway being coregulated by an operator located between the *bioA* and the *bioBFCD* genes (Otsuka and Abelson 1978). Frequently one, or both, of the divergent genes encodes a regulatory protein which acts directly on the bidirectional promoter. This is facilitated in prokaryotes where translation is coupled to transcription so that there is a high concentration of the regulatory protein near to its place of transcription. The *cI* and *cro* genes in bacteriophage lambda are an example of two regulatory genes with a central promoter control region which acts as a genetic switch so that one or other of the two genes is expressed (Ptashne 1986).

Tandemly arranged promoters occur where there are adjacent genes or operons and also where there is more than one promoter directing transcription of a gene or part of an operon. This arrangement of promoters is very common in prokaryotes and may cause promoter occlusion of the downstream promoter, but it is not clear to what extent this arrangement is important for gene control. Transcription from the bacteriophage lambda P_L promoter through a downstream, tandemly arranged, promoter has been shown to inhibit the activity of the downstream promoter (Adhya & Gottesman 1982). In this case transcription from the very strong P_L promoter was shown to inhibit transcription from the two *gal* operon promoters, P_{g1} and P_{g2}. In some cases differential efficiency of translation

from mRNAs derived from different promoters may be part of gene regulation by promoter occlusion (Katzir et al 1976).

Transcription through a promoter from another convergently arranged promoter may also cause inhibition of promoter activity. Expression from the *trp* promoter, measured by polypeptide synthesis, is reduced when there is convergent transcription from the lambda promoter P_L (Ward & Murray 1979). However, this inhibition may critically depend on the relative levels of transcription from the two promoters, as expression from the *trp* promoter, measured by RNA synthesis, is not affected significantly when it is faced by the *lacUV5* promoter (Horowitz & Platt 1982). There are three known prokaryotic promoters which are close together and transcribe convergently across each other, but do not form a significant length of antisense RNA, and the regulatory implications in these cases is not known (Beck & Warren 1988).

Apart from promoter inhibition, convergent promoters allow the formation of antisense RNA. Antisense regulation has been demonstrated in a wide variety of prokaryotic systems (Green et al 1986). Negative control of the copy number of ColE1 type plasmids is accomplished by RNA I which is transcribed in the opposite direction from the same DNA that encodes the primer RNA for DNA replication at the origin (Polisky 1988). The transposase of the insertion sequence IS10 is negatively controlled by a small complementary RNA specified by the IS10 pOUT promoter (Simons & Kleckner 1983). The *E. coli* outer membrane porin gene, *OmpF* is negatively regulated by a 174 bp RNA molecule called *micF* (Mizuno et al 1984). *micF* is 70% identical to the inverse of the 5' region of the *OmpF* mRNA but is transcribed from another region of the genome. Indeed, *micF* is co-regulated with *OmpC* via a bidirectional promoter and OmpC is the other major membrane porin. The cAMP receptor protein (CRP) gene is autoregulated by the cAMP-CRP complex inducing the synthesis of an

antisense RNA (Okamoto & Freundlich 1986). The transcription start sites of the antisense RNA and the *crp* mRNA are overlapping, 3 bp apart and on opposite DNA strands. Although the DNA specifying the anti-sense RNA and the DNA specifying the *crp* mRNA are separate sequences, the RNAs are complementary. Other examples are found in bacteriophage lambda, bacteriophage p22 and the plasmid pT181.

The bacterial chromosome is divided into topologically independent domains. Furthermore, the degree of supercoiling can influence promoter strength, so it is feasible that domains of genes are coregulated by controlling the supercoiling of the domains (Pruss & Drlica 1989). However there is no evidence of clustering of functionally related operons in the *E. coli* genome and domains of gene regulation have not been demonstrated.

In general bacteria utilize a wide variety of regulatory strategies which conserve the amount of regulatory DNA and rely on the *cis* location of genes. It is of interest to know whether these regulatory strategies are also widely used in the mammalian genome.

Do these mechanisms function in eukaryotic cells

The prokaryotic genome has been studied in great detail and there are numerous examples of sharing of regulatory elements and interaction between neighbouring genes. There are very few such examples amoung eukaryotic genes and this could simply be due to the fact that the eukaryotic genome has been less studied. Alternatively it is possible that the transcription machinery in eukaryotic cells is different from prokaryotic cells so that sharing of regulatory elements, promoter interference and antisense inhibition cannot occur.

Regulatory elements can be shared as shown by the use of bidirectional promoters. Bidirectional promoters are found in the papovavirus' where they act as a genetic switch between the early and late genes (Tooze 1981).

Bidirectional promoters are also found in yeast where they allow coregulation of the flanking genes by a centrally placed activator sequence (West et al 1984). In addition eukaryotic genes are regulated by "enhancers" which are binding sites for trans-acting transcription activation proteins which can act at a distance (Serfling et al 1985). Enhancers can act at a distance, in an orientation independent manner and from either upstream or downstream of a gene thus allowing an enhancer to be shared by more than one gene. If an enhancer is located to one side of two genes, it preferentially stimulates the nearer promoter, but it does affect both genes (A. Lennard personal communication).

Promoter occlusion between tandemly arranged promoters has been shown to occur. In retrovirus' only the 5' promoter is active, however if the 5' promoter is inactivated or deleted, then the 3' promoter is found to be an efficient promoter (Cullen et al 1984). In an experimental construct containing two α-globin genes, the downstream promoter is inhibited unless a terminator of transcription is placed between the two genes or the upstream gene is inactivated (Proudfoot 1986). The drosophila alcohol dehydrogenase gene is transcribed from two closely linked promoters (Corbin & Maniatis 1989). The proximal promoter is active primarily in first to early third-instar larvae, whereas the distal promoter is active in late third-instar larvae and adults. The switch from the proximal to the distal promoter is caused by the up regulation of the distal promoter which shuts off the proximal promoter by promoter occlusion.

Inhibition of gene expression by the presence of antisense RNA transcribed from the complementary DNA strand has been shown to operate experimentally, but has not been observed as a natural regulatory mechanism. Antisense control of translation by small control RNAs (which are not necessarily transcribed from the same portion of the genome as the mRNA) has been suggested to occur (Heywood 1986). The 3' end of short

translation control RNAs, called tcRNA102, is postulated to bind to the 5' end of specific muscle mRNAs and cause translational control in the same way that other snRNAs bind at splice sites and control splicing.

There are numerous examples of experimental systems where over expression of an antisense gene gives significant inhibition of a nuclear gene (van der Krol et al 1988). For example transfection of antisense HSV-TK into mouse L cells stably transformed with HSV-TK gives a six fold reduction in TK activity (Izant & Weintraub 1985). However it has often proven difficult to achieve complete blocking of gene expression by inhibition with antisense RNA, often large excess amounts of antisense RNA are needed to see phenotypic effects, and sometimes it has not been possible to get any effect at all (van der Krol et al 1988). There is now evidence for the widespread expression of a double stranded RNA unwindase which is present in Xenopus embryos but not in frog oocytes (Bass & Weintraub 1987, Rebagliati & Melton 1987). Unwindase separates double stranded RNA into two single stranded forms which are modified in some way but it is not clear whether this single stranded RNA is still in a translatable form (Bass & Wientraub 1988). Unwindase activity is present in a variety of mammalian cell lines including HeLa, human lymphoblastoid, mouse plasmacytoma, Burkitt's lymphoma and mouse F9 teratocarcinoma (Wagner & Nishikura 1988). In addition unwindase activity is very low in mouse fibroblast 3T3 cells arrested into quiescence but increases when the cells are stimulated by fetal calf serum (Wagner & Nishikura 1988). If the lack of unwindase activity in Xenopus oocytes is responsible for the observation that antisense RNA causes inhibition in Xenopus oocytes but not in Xenopus embryos, then the very variable levels of inhibition observed with antisense RNA in mammalian cells may be due to variable amounts of unwindase being present in these cells.

Clearly regulatory elements can be shared in mammalian cells, transcription from an upstream promoter through a downstream promoter can cause promoter occlusion and antisense RNA can cause antisense inhibition.

1.2 Organization of the Mammalian Genome.

The Mammalian genome has been studied in far less detail than the prokaryotic genome and although the molecular mechanisms exist for the sharing of regulatory elements and the interaction of neighbouring transcription machinery, it is not clear to what extent these mechanisms are used in gene regulation. The drosophila genome is considerably smaller than the mammalian genome and has been studied in great detail at the molecular level. The transcription factors in all eukaryotes have been found to be remarkable conserved, to the extent that the yeast TATA-box binding transcription factor TFIID substitutes for the mammalian equivalent in an *in vitro* transcription system (Buratowski et al 1988). As transcription factors are highly conserved it is expected that regulatory mechanisms will also be highly conserved, so examples from the drosophila genome are used in the following discussion.

Cistron like arrangements

Eukaryotic genes are not arranged together on cistronic mRNA, but there are three mechanisms which achieve a similar regulation of more than one gene from a single promoter. The three large rRNAs, 18S, 5.8S and 28S are initially transcribed as a long precursor RNA and this is then processed to give the three individual rRNAs (John & Miklos 1988 p179). Multiple different mRNAs are sometimes made from a single precursor RNA by differential splicing, and the resulting poly(A)+ mRNAs encode different proteins (Leff & Rosenfeld 1986). Differential splicing allows an additional

level of control to be exercised over the relative amounts of the various proteins made. Finally, joining of different repeated sections of the immunoglobulin genes allows the production of an enormous variety of antibodies from a single parental gene (John & Miklos 1988 p 100)

The mammalian genome has one other major method of regulating expression of genes which is dependent on the position of the gene in the genome. This is the inactivation of one of the two female X chromosomes so that the genes on that chromosome are not expressed. X inactivation is thought to be correlated with hypermethylation of the DNA and is dependent on the X inactivation centre (Lyon 1988a). The nature of the X inactivation centre is not known.

Genes are individually regulated

In general mammalian genes are independently regulated at the level of transcription (Dynan & Tjian 1985). Transcription is controlled by DNA sequences located near the gene which bind various transcription factors. The main evidence for this has come from cloning and reintroducing genes into cells either transiently or stably by $CaPO_4$ precipitation. This shows that in general genes are correctly regulated if they have several kilobases of flanking DNA on either side. Cloned genes can also be introduced into whole animals by injecting the DNA into fertilized eggs or by making chimeric mice with altered embryonal stem cells. Again this shows that one can get correct differential regulation of a gene if several kilobases of flanking DNA are included (Palmiter & Brinster 1986). Stable transfectants and transgenic mice do however show that genes at different chromosomal locations are not equally expressed due to the different sites of integration (Palmiter & Brinster 1986).

When a gene integrates into heterochromatin it is usually inactivated and this is a reflection of the chromatin conformation of the DNA.

Heterochromatin consists almost exclusively of highly repetitive satellite DNA. The best studied satellite DNA is the human alpha satellite which has been shown to consist of tandem repeats of a 171 bp unit covering hundreds of kilobases of DNA (Waye et al 1987). There is no evidence of genes being present in the heterochromatin and there is no evidence of this type of inactivation being used as a means of gene control. The exception is X inactivation where the whole chromosome becomes heterochromatic.

Position effects on transfected DNA consist of an overall lowering of transcription and expression of the transfected gene in tissues in which it would not normally be expressed (Palmiter and Brinster 1986). This indicates that although the DNA located within several kilobases of a gene specifies differential regulation, it often does not include all the DNA needed to screen the gene from the effects of nearby chromatin conformation and nearby transcription regulatory elements. Recently sequences have been isolated which allow position independent transcription of genes introduced into transgenic mice. A DNA fragment of 6.5 kb from the upstream side of the human β-globin cluster allows position independent expression of β-globin in transgenic mice (Talbot et al 1989). This dominant activating sequence is normally located some 50 kb away from the β-globin gene, beyond the ε, Gγ and Aγ and δ genes. Similarly a 5.5 kb DNA fragment conferres efficient position independent expression of a linked human CD2 minigene in transgenic mice (Greaves et al 1989). The dominant activating sequence contains a hypersensitive site which is located 500-1000 bp downstream of the 3' end of the CD2 gene. Both the CD2 and β-globin dominant activating sequences can operate when placed in either orientation with respect to the transcription of the responder gene, and when placed upstream or downstream of the responder gene. In addition, the CD2 dominant activator sequence confers position independent expression on the human β-globin gene in transgenic mice such that the β-globin gene is fully expressed in the

thymus and not expressed in the brain. However, the CD2 dominant activator sequence does not confer position independent transcription of the β-globin gene in erythroid tissues. In erythroid tissues expression of the β-globin gene is position-dependent indicating that the CD2 dominant activator sequence is tissue specific (Greaves et al 1989).

Dominant activating sequences have so far been found at some distance from the regulated gene or 3' to the gene. Most transgenes have been made with only a limited amount of DNA from the 5' side of the gene and probably have not contained a dominant activating sequence, and the result has been position effects on gene expression. It remains to be seen whether all genes have an associated dominant activator sequence which isolates them from the transcriptional state of the surrounding DNA.

Highly repetitive dispersed DNA

A major feature of the mammalian genome is the enormous number of dispersed repetitive elements. In the mouse these repetitive elements fall into two classes, short interspersed elements (SINES) and long interspersed elements (LINES), neither of which have been shown to have any function in the genome.

There are two families of mouse SINEs, the B1 and the B2 elements. B1 and B2 elements are structurally similar being 129 and 180 bp in length respectively, flanked by direct repeats of 9-18 bp and with an A rich region at the 3' end (Rogers 1985). It is thought that they have spread throughout the genome by retroposition (Rogers 1985). B1 sequences differ by 15% on average from the consensus sequence and B2 elements differ by 8.8% indicating that both families have arisen recently in the mouse genome (Rogers 1985). One B1 element which is actively transcribed from its internal polIII promoter has been found (Young et al 1982) but most are

inactive. B2 transcripts are abundant in cells but it is not clear how many elements are actively transcribed.

Many mammalian genes contain SINES in their introns or untranslated regions and these might be expected to interfere with the expression of the gene. If the element is expressed itself then promoter occlusion and antisense inhibition would be expected to occur. Even if the element is not expressed, these elements are highly represented in the mRNA population due to co-expression with genes into which they have integrated, so one would expect to get antisense interference. One example of a B1 element located in the first intron of mouse *α-fetoprotein* gene is reported by Young et al (1982). When the mouse *α-fetoprotein* gene is transfected into African green monkey kidney cells on an SV40 vector, both the *α-fetoprotein* and the B1 mRNAs are expressed and can be detected by S1 mapping. Apparently the two genes function quite effectively. Even though there has been a vast increase in the number of elements in the mammalian genome over the last few million years, gene regulation seems little affected indicating that the mammalian genome is largely resistant to interference by nearby promoters and antisense RNA due to SINES.

The most abundant long interspersed elements (LINEs) in mice are a single family of elements called L1. There are about 100,000 copies per mouse genome of which a large number are truncated to a variable degree at the 5' end, but the full length copies are about 7 kb long (Casavant et al 1988). It is thought that L1 elements are spread by retroposition as they have A rich tracts at the 3' end and generate short duplications on integration. At least some L1 elements are actively transcribed in human undifferentiated teratocarcinoma cells (Skowronsik & Singer 1986) but truncated elements are probably not actively transcribed. L1 elements are distributed throughout mammalian species but are more similar within a species than between species. Comparison of the distribution of L1 elements

in the β-globin loci of *Mus caroli* and *M. domesticus* has shown that they have dispersed in the mouse genome within the last 2 million years (Casavant et al 1988).

Insertion of L1 elements can cause incorrect gene expression. In 2 out of 240 unrelated patients with hemophilia A, the mutation was found to be the insertion of an L1 element which was not present in the parental DNA (Stylianos & Kazazian 1988). In both cases the L1 element had integrated into intron 14, but they were integrated in different orientations in relation to the *factor VIII* gene. It is not known how the L1 elements are disrupting *factor VIII* expression.

SINES and LINES have proliferated in the mammalian genome to a remarkable degree and usually have little affect on gene expression. Indeed the lack of effect on interrupted genes might be a reason why SINEs and LINEs have been able to accumulate to such a degree without causing a lethal mutation load. Retrovirus like inserts and some Drosophila transposable elements, on the other hand, are frequently found associated with incorrectly expressed genes and it is the transcription machinery of the virus which is often responsible for the mutation.

A classic example of a mutation caused by a viral insertion is the dilute (*d*) coat colour mutation of DBA/2J mice which is associated with the integration of an ecotropic MuLV genome (Jenkins et al 1981).

Viral insertions have been implicated in oncogenesis by activating proto-oncogenes (Varmus 1984). In mice, MMTV has been found integrated nearby two genes, *int1* and *int2* in the mammary tumors which the virus is thought to cause. A study of the viral integrants near the activated *int1* and *int2* genes has shown that the virus is normally integrated on either side of the activated gene and is transcribed away from the gene (Nusse 1988). It is thought that the viral enhancer in the LTR is causing activation of the *int1* and *int2* genes.

In drosophila eight mutations at the *Notch* locus are caused by integrations of copia like transposable-elements (Kidd & Young 1986). All the transposable elements are integrated in the first or second intron of the *Notch* gene and are transcribed in the reverse direction. These mutations could be due to the integrated elements disrupting *Notch* regulatory elements in the introns. However, as the type of transposable element and the orientation of transcription of the element seems to be a feature of the mutations, it is thought that the transcriptional machinery of the element is interfering with transcription of the *Notch* gene.

The insertion of SINEs and LINEs into genes can cause incorrect gene expression but considering the vast number of these elements they affect gene expression very little. This indicates that the presence of antisense RNA and transcriptionally active promoters inside genes need not cause inhibition of gene expression. Viral elements on the other hand are very mutagenic and the regulatory elements of the viruses are often found to have interacted with the endogenous genes transcriptional machinery.

CpG-rich islands

A striking feature of mammalian genomes which is completely lacking in drosophila and yeast is the occurrence of CpG-rich islands. The bulk nuclear DNA of mammals contains about 41% G and C, and the dinucleotide CpG occurs at about one fifth of the expected frequency (Russell et al 1976). Some 60-90% of CpG dinucleotides are methylated at the 5 position on the cytosine ring and this accounts for about 50% of the total methylcytosine in the mammalian genome (Woodcock et al 1988). The low level of CpG dinucleotides is probably caused by deamination of 5-methylcytosine to give thymine. When this is not repaired correctly it results in a CpG dinucleotide being mutated to TpG. About 1% of vertebrate DNA has been found to have completely the reverse properties to the bulk. This DNA is located in CpG-

rich islands. These islands are stretches of DNA containing about 65% G+C, show no suppression of the CpG dinucleotide and are not methylated in any tissues except on the inactive X chromosome (Bird et al 1985, Bird 1986). It is possible to clone DNA from these CpG-rich islands by cloning small fragments of genomic DNA cut with the enzyme HpaII which has the recognition sequence CCGG and is methylation sensitive (Bird et al 1985).

Two lines of evidence lead to the conclusion that CpG-rich islands are associated with the 5' ends of genes. Firstly, islands which have been cloned out randomly have been found to be associated with genes (Lindsay & Bird 1987, Lavia et al 1987). Indeed this has been used to clone a gene in the t-complex (Rappold et al 1987) and a candidate gene for the cystic fibrosis locus (Estivill et al 1987). Secondly, genes which have been cloned by other means have been found to have CpG-rich islands at their 5' ends. A recent survey of RNA polymerase II transcribed genes in the GenBank DNA sequence data bank (Gardiner-Garden & Frommer 1987) has found CpG-rich islands associated with approximately two thirds of the genes. However this sample of genes is extremely biased due to the ease of cloning highly expressed, tissue specific genes in comparison to low expressed and "housekeeping" genes, thus giving an underestimate of the proportion of all genes with associated CpG-rich islands. In the survey, all ubiquitously expressed genes have an associated CpG-rich island suggesting that as the proportion of cloned genes which are "housekeeping" genes grows, so will the proportion of cloned genes with CpG-rich islands.

From the fraction of the vertebrate genome which is sensitive to the enzyme HpaII it has been calculated that there are some 30,000 islands with an average spacing of 100 kb (Bird et al 1985). Estimates of the number of distinct mammalian RNA messages are variable. Numbers are derived from measuring RNA complexity (Chaudhari & Hahn 1983) or from the number of lethal complementation groups in organisms such as drosophila (Miklos &

John 1988 p146). Both methods have drawbacks, RNA complexity doesn't take into account differential splicing and the presence of repetitive elements in mRNAs from unique genes. Counting lethal complementation groups misses out large numbers of genes which occur in more than one copy and genes which are inessential (Ohno 1985). Inessential genes in drosophila include the metabolic enzymes esterase-C, esterase-6, alkaline phosphatase, acid phosphatase, xanthine dehydrogenase, aldehyde oxidase, alcohol dehydrogenase, isocitrate dehydrogenase, α-glycerophosphate dehydrogenase, tryptophan pyrolase and kynurenine hydroxylase (John & Miklos 1988 p328). Brain tissue is by far the most complex tissue in the mammal, and from RNA complexity has been calculated to have about 30,000 housekeeping and 124,000 brain specific active genes (Bernadi et al 1985). The nature of the very large number of brain specific genes is not known but the number of housekeeping genes corresponds well to the number of CpG-rich islands.

The presence of CpG-rich islands is a powerful way of detecting the presence of genes in genomic DNA. Although tissue specific genes may be missed, the only other methods of detecting genes in large stretches of DNA rely on hybridization with specific probes generated from cloned DNA, which is laborious. One such method to detect genes is to look for DNA conserved between species, and this has been used to locate the testis-determining factor gene in 230 kb of cloned DNA (Page et al 1987) but this method is liable to miss any genes which are evolving rapidly and will be sensitive to the detection of pseudogenes of highly conserved genes. Alternatively hybridization of probes to Northern blots detects transcribed DNA. This method will miss genes which are transcribed only in a few tissues unless Northern blots of all differentiated tissues and developmental stages are used. It also requires the DNA probes to be unique DNA, as repetitive sequences occur in mRNA. Competitive hybridization can be used to

overcome repetitive sequences but this reduces the sensitivity so that messages expressed at a very low level will not be detected.

CpG-rich islands are characterized by their C and G base composition and lack of methylation, and this in turn gives them a characteristic set of restriction enzyme cleavage sites. Enzyme recognition sites which consist of six C or G nucleotides occur very rarely in mammalian genomic DNA because of the low average C and G content and these sites are concentrated in CpG-rich islands. Sites which consist of eight C or G nucleotides are considerably rarer because of the increased size of the recognition site. If the recognition site contains one or more CpG dinucleotides then it is further confined to CpG-rich islands due to the reduced representation of CpG in non-island DNA. Furthermore a number of enzymes which have CpG in their recognition sites are sensitive to methylation (Nelson and McLelland 1987), so even if a site does occur outside a CpG-rich island it will probably be methylated and will not be cut by the enzyme. Table 1.1 shows a variety of rare cutting enzymes and the degree to which their cutting sites are restricted to islands. For localizing islands by southern blotting the most useful enzymes, in other words, those which are restricted to islands to the greatest degree but are not too rare, are; SacII, BssHII, EagI, SmaI, NaeI, and NarI.

Table 1.1 Island specific rare cutting enzymes.

Enzyme	Site	CpGs	sensitivity to 5 mC	% sites in islands	Av. size of frags.
NotI	GCGGCCGC	2	+	89	1,000 kb
EagI	CGGCCG	2	+	74	100
SacII	CCGCGG	2	+	74	100
BssHII	GCGCGC	2	+	74	100
SmaI	CCCGGG	1	+	42	100
NaeI	GCCGGC	1	+	42	100
NarI	GGCGCC	1	+	42	100
MluI	ACGCGT	2	+	27	300
PvuI	CGATCG	2	+	27	300

data from Lindsay and Bird 1987 and Biolabs catalog 1988

Using the presence of CpG-rich islands to map genes

Although CpG-rich islands do not occur at the 5' ends of all genes, detection of CpG-rich islands is an extremely powerful method of locating genes which do have islands. Long range mapping of the genome with island specific rare cutting enzymes determines the average spacing of islands and hence genes in these regions and indicates how closely genes are spaced.

Long range mapping with island specific rare cutting enzymes has been carried out around two randomly cloned CpG-rich islands and the island at the 5' end of the mouse dihydrofolate reductase gene (*dhfr*) (Brown & Bird 1986). These studies showed an average spacing of 109 kb between neighbouring islands, with the spacing ranging from 40 to 400 kb.

The entire region of 10 mega base pairs (Mb) surrounding the Duchenne Muscular Dystophy gene (DMD) has been mapped with rare

cutting enzymes (Burmeister et al 1988). At the proposed 5' end of the DMD gene there is a CpG rich region spanning about 300 kb and containing NaeI, NruI, SacII and BssHII sites. However, this does not correspond to a CpG-rich island as all the sites are methylated in blood DNA and there is no CpG-rich island detectable in the published sequence of the 5' end of the gene. There are no other recognizable islands within about 1 Mb on either side of this CpG rich region, indeed the mapping was carried out using sperm DNA as all the rare cutting enzyme sites were methylated in blood DNA. The region around the DMD gene is thus completely devoid of CpG-rich islands for at least 2 Mb, though this does not rule out the presence of other tissue specific genes which do not have CpG-rich islands.

The region proximal to the *MIC2* gene in the pseudoautosomal region has been mapped with island specific rare cutting enzymes (Pritchard et al 1987). This study revealed that on the X chromosome there is an island ~150 kb proximal to the *MIC2* island and on the Y chromosome there is an island ~260 kb proximal to *MIC2*.

The pseudoautosomal region of the human X and Y chromosomes has been mapped for rare enzyme cutting sites (Petit et al 1988, Rappold & Lehrach 1988, Brown 1988). This shows that over the entire 2000 kb of this region, excluding the telomeric 200 kb, at least 7 CpG-rich islands occur roughly evenly distributed. The spacing varies from 100 to 400 kb which is consistent with other estimates of average spacing between CpG-rich islands. The telomeric 200 kb however is very C+G rich and contains a lot of cutting sites for island specific rare cutting enzymes. However this clustering of unmethylated CpG dinucleotides does not represent a dense clustering of CpG-rich islands (Petit et al 1988).

Regions round a series of DNA markers near to the Huntington's disease gene on human Chromosome 4p15-4pter have ben mapped with NotI (Pohl et al 1988). One region covering three probes spans 175 kb and is very rich in

unmethylated CpG sequences, containing seven NotI sites with 10-70 kb between adjacent sites. Considering that only one tenth of CpG-rich islands are expected to contain a Not I site (table 1.1) this could indicate the presence of 70 CpG-rich islands in this 170 kb !. Although this is a telomeric region, the high concentration of CpG is thought to represent CpG-rich islands rather than a general high C+G content as was observed at the telomeric region of the pseudoautosomal region (Petit et al 1988). Probes taken from CpG-rich regions of 4p15-4pter show conservation between species indicating that they probably represent genes (Pohl et al 1988).

The CpG-rich islands in the α-globin cluster have been mapped showing that here four CpG-rich island are located in 30 kb of DNA (Fischel-Ghodsian et al 1987).

There are numerous other long distance mapping studies but these are on the whole not informative for the presence of CpG-rich islands as they use a variety of rare cutting enzymes whose cleavage sites are not CpG-rich island specific.

The mapping studies with island specific rare cutting enzymes indicate that CpG-rich islands are generally well spread out over the genome with an estimated average spacing of 100 kb. However, there are regions of the genome which are densely covered with CpG-rich islands and others which are completely lacking islands. Furthermore a cluster of very closely spaced CpG-rich islands would probably appear as a single island in some of these long range mapping studies.

Bidirectional promoters in CpG-rich islands

It is becoming a commonly observed feature of CpG-rich islands that they contain bidirectional promoters. In general the two genes associated with the bidirectional promoter are not related by DNA sequence and are not

necessarily coregulated. No detailed molecular analysis of the transcription factors involved has yet been carried out.

One randomly cloned mouse CpG-rich island contains the 5' ends of two poly(A)+ RNA transcripts (Lavia et al 1987). To the left is a 0.7 to 1.0 kb mRNA and to the right is a 2.7 kb mRNA. The cDNAs of both genes contain long open reading frames and both genes also contain introns, but there is no sequence identity between them. The 5' end of the two genes were mapped by S1 nuclease mapping and shown to be coincident and divergent. However there were multiple start sites in both directions which is consistent with the observed lack of any TATA boxes. Both genes are expressed in liver RNA and embryonic mouse RNA indicating that they may be ubiquitously expressed genes though expression in different tissues was not directly addressed.

The Chinese hamster *dhfr* gene is transcribed from a promoter located in a CpG-rich island. At 132 bp upstream of the major start site of transcription of the *dhfr* gene is the start site of a divergently transcribed gene (Mitchell et al 1986). The divergent gene specifies a polyadenylated RNA transcript of a 4 kb the abundance of which increased threefold after serum starvation of Chinese hamster ovary (CHO) cells for 24 hrs. Under similar conditions the level of *dhfr* mRNA does not change. The two genes do not share sequence identity.

A genomic segment from African green monkey has been cloned by homology to a regulatory region of SV40 (Saffer & Singer 1984). This segment is a CpG-rich island as shown by its CpG content but is not unique being a member of a family consisting of about 80 members. Both strands of the cloned DNA hybridize to cellular transcripts and the cloned DNA directs bidirectional transcription when inserted into an expression vector and transfected into monkey cells. S1 nuclease experiments indicate that divergent transcription is initiated from sites 212 bp apart but again there

are multiple start sites for both directions of transcription and no TATA boxes. Nothing is known about the sequence of the putative flanking genes or their differential expression.

There are a number of reports of bidirectional transcription from promoters placed in expression vectors or otherwise removed from their normal environments. However these promoters do not direct transcription of two divergent genes in their normal locations.

The Rat Insulin II 5' flanking region directs transcription exclusively in the β cells of the endocrine pancreas. Transgenic mice containing this promoter develop insulinomas due to transcription of the SV40 large T antigen (Efrat & Hanahan 1987). Transcription is directed away from the normal insulin II gene, starting at -255, and is specific for β cells of the endocrine pancreas. However no stable transcript corresponding to this region upstream of the insulin gene could be detected in insulinoma RNA, and the DNA sequence of this region contains no open reading frames.

The chicken skeletal α-actin gene promoter region directs transcription specifically in myotubes. When 413 bp of this promoter are placed in front of CAT and transfected into chicken primary myoblasts, it directs increased transcription as the myoblasts differentiate into myotubes. When placed in the reverse orientation, the same promoter fragment directs transcription of the test gene with the same differential regulation. (Grichnik et al 1988). However, divergent transcripts are not detectable from the endogenous skeletal α-actin gene by RNA dot-blot assay, and S1 nuclease analysis of RNA from 15 day embryonic cardiac breast tissue, brain tissue and muscle culture tissue failed to detect any protected fragments.

The mouse c-Ki-*ras* gene is ubiquitously expressed from a promoter which is located in a CpG-rich island. No genes have been detected directly upstream from c-Ki-*ras*. However, a 688 bp SmaI fragment containing the promoter region directs transcription equally well when placed in either

orientation in front of a CAT gene which is transfected into CV-1 cells (Hoffman et al 1987).

It is clear from the above that promoters in CpG-rich islands are often capable of directing bidirectional transcription, however, the majority do not direct transcription of two divergent genes. But there is a class of bidirectional promoters located at CpG-rich islands driving transcription of two genes. The Chinese hamster *dhfr* gene promoter is the only promoter which directs the transcription of two unrelated divergent genes where the differential regulation of these genes has been analysed. Although both genes may share transcription factors directing ubiquitous basal levels of transcription, the two genes are not coregulated.

Clusters of repeated genes.

A major feature of evolution of the mammalian genome is duplication of genes and recruitment of the new genes for different purposes. This process is very obvious in the mammalian genome where the duplicated genes are located in clusters. Obvious examples are the α-globin and β-globin clusters but it is a very wide spread occurrence, there being clusters of enzyme genes (Stevenson et al 1986), growth factors receptors (Roberts et al 1988) and homeobox genes (Colberg-Poley 1985) amongst many others.

When individual genes from the human β-globin cluster are introduced into transgenic mice there is tissue and developmental stage specific expression of the genes (Kollias et al 1986). The human fetal Aγ-globin gene is expressed like an embryonic mouse globin gene and the human β-globin gene is expressed like an adult mouse globin gene. In the case of the Aγ-globin gene 3.3 kb of DNA containing the globin gene, 1.3 kb of 5' flanking DNA and 370 bp of 3' flanking DNA is enough to specify correct differential expression. In the case of the β-globin gene the gene with 0.8 kb of 5' flanking and 2.5 kb of 3' flanking DNA gives correct expression. This shows

that these two genes in the β-globin cluster have their own regulatory elements which are enough to give tissue and developmentally controlled expression. However, there is now evidence for controlling elements which affect the whole cluster (see section "Domains containing coregulated genes"). The β-globin cluster is one of the best studied regions of the mammalian genome and the regions controlling the whole domain have only recently been identified. It may be that other clusters of duplicated genes are also controlled by regulatory elements which affect the whole cluster, the long range regulatory elements have not yet been found and individual genes are independently regulated like the β-globin genes.

An exception to the independence of clustered duplicated genes is found in the chicken β-globin cluster. A single enhancer is required for the erythroid specific promoter activity of both the β- and ε-globin genes when these genes are separately transfected into erythroid cells. The enhancer is located downstream of the β-globin gene, about 1.6 kb 3' to the β promoter, and upstream of the tandemly arranged ε-globin gene, about 1.5 kb 5' to the ε-promoter (Nickol & Felsenfeld 1988). When the adult β-globin gene and enhancer are transfected into erythroid cells of different developmental stages, there is correct developmental expression. However the embryonic ε-globin gene with the enhancer is expressed in both embryonic and adult red blood cells. When plasmids containing both the β- and ε-globin genes and the shared enhancer are used, the ε-globin has correct cell type and stage specific expression indicating that the β-globin gene is needed for correct regulation of the ε-globin gene (Choi & Engel 1988). Detailed analysis indicates that there is a Stage Selector Element (SSE) within the adult β-globin promoter which preferentially interacts with the cell type specific enhancer. In embryonic cells, the β-globin SSE is inactive and the ε-globin gene interacts with the shared cell type specific enhancer and is expressed. In adult erythroid cells the β-globin SSE is active and the β-globin

promoter preferentially interacts with the shared enhancer, switching off the ε-globin gene (Choi & Engel 1988)

A slightly different situation is found with the mouse type IV collagen genes. The α1(IV) and α2(IV) collagen genes are arranged head to head with the 5' ends separated by 130 bp. Type IV collagen is a heterotrimer composed of two α1(IV) and one α2(IV) chains. The two genes are coregulated being expressed at low levels in undifferentiated F9 teratocarcinoma cells and induced in large amounts with retinoic acid. The two genes share a bidirectional promoter and there is an enhancer within the first intron of the α1(IV) collagen gene which directs efficient transcription of both genes (Burbelo et al 1988). Both genes are collagen genes indicating that they arose by duplication but the homology between them is weak (Poschl et al 1988).

The human complement protein Factor B is encoded by a single gene closely linked to the gene encoding the second component of complement C2. The poly(A) addition site of C2 is located just 421 bp from the transcription initiation site of the Factor B gene (Wu et al 1987). Deletion analysis of the promoter of Factor B shows that DNA located in the 3' end of the C2 gene is important for the transcription of Factor B. These two genes are related by structure as well as function suggesting that they arose by duplication (Carroll et al 1985). It is not clear whether Factor B and C2 share regulatory elements, but it is clear that promoter occlusion between the two tandemly arranged promoters does not lead to alternate expression

The rat c-*erbA*α*-2* gene overlaps the gene for a related thyroid/steroid hormone receptor (Lazar et al 1989). The genes are assumed to have arisen by duplication and overlap at their 3' ends by 269 bp. Antisense interference between the genes has not been investigated. There is also the possibility of the overlap affecting splicing of the c-*erbA*α gene as there are two differently spliced forms of this gene, *c-erbA*α*-2* which overlaps the

convergent gene, and *c-erbAα-1* which does not (Lazar et al 1989). Rat *c-erbAα-1* is a nuclear receptor protein which binds thyroid hormone (T3) and regulates gene expression by binding to specific DNA sequences near hormone-responsive genes (Koenig et al 1989). The *c-erbAα-2* protein does not bind T3 and inhibits the T3-dependent inductive effect of both *c-erbAβ* and *c-erbAα-1* on expression of a T3-responsive test gene (Koenig et al 1989). It is possible that the antisense RNA from the convergently transcribed gene is affecting the splicing of *c-erbAα* and thus changing the ability of the cell to respond to T3.

Gene duplication has given rise to some very large complex arrangements of genes such as the major histocompatability complex. Where the individual genes are part of a very rapidly evolving biological system such as the immune system, the genes have sometimes lost almost all signs of homology at the DNA level and are only related by tertiary protein structure. This can give rise to clusters of genes which are clearly all part of a single biological unit, and are coregulated, but are apparently unrelated at the DNA or protein sequence level. However this should not necessarily be taken as evidence of coregulated domains of DNA. Although the genes are no longer homologous at the DNA level they arose quite recently in evolution. An example of this is the mouse IL-3, IL-4, IL-5 and GM-CSF genes which are clustered on mouse chromosome 11 and human chromosome 5 (Takahashi et al 1989). Although these genes are not related at the DNA level they are probably related phylogenetically.

In general, the clustering of replicated genes is a result of the duplication process (Maeda & Smithies 1986) and is not important for coregulation of the clustered genes. Indeed the clustered genes often have different tissue specificities. However the above examples of shared transcription elements and interacting transcription machineries indicate

ways in which the clustering of duplicated genes has been exploited towards gene regulation.

Clusters of unrelated genes in mammalian genomes

The mammalian genome is so large and so few genes have been cloned that it is perhaps not surprising that very few cases of close but unrelated genes have been found. The number of such close genes is however growing rapidly and it is becoming possible to assess whether there is any functional significance to their clustering. Genes located on either side of bidirectional promoters have already been described (see section "Bidirectional promoters in CpG-rich islands")

A gene called *D15Kz1* has been detected just 3.4 kb upstream of the mouse Glycerol-3-phosphate dehydrogenase gene (*Gdc-1*) (Johnston et al 1989). *D15Kz1* directs transcription of two mRNAs of 3.2 and 3.4 kb in length, transcribed in the same direction as *Gdc-1*. The two genes have no DNA or protein sequence similarity. Although both genes are ubiquitously expressed, they are both differentially regulated in different tissues, and the differential expression of the two genes is completely independent.

An unrelated gene has been found just upstream of the *fes/fps* proto-oncogene (Roebroek et al 1986). The 3' end of this upstream gene, called *furin*, lies less than 1.1 kb upstream of the 5' end of the *fes/fps* gene. The two genes are not related by DNA or protein sequence but they might be functionally related. As the promoter region of *fes/fps* has not yet been identified, interaction between the two genes cannot be evaluated.

The rat gonadotropin-releasing hormone gene (*GnRH*) overlaps with a convergently transcribed gene called *SH* (Adelman et al 1987). The two genes share exon sequences but they are not expressed in the same tissues. *GnRH* is expressed in the central nervous system while *SH* transcripts are present in the heart. The *SH* gene has not been completely sequenced but it

does not appear to be related to *GnRH* by DNA sequence or function. As the genes are not expressed in the same tissues there is no reason to think that anti-sense inhibition is occurring.

The ubiquitously expressed mammalian histone genes are clustered together in small groups (Marzluff & Graves 1984). Thus individual lambda clones of mouse DNA contain up to four histone genes, belonging to the various unrelated histone gene classes. The order of the histone genes in different λ clones is not the same and appears quite random. It is interesting to note that where there are isolated histone genes, these have become tissue specific and their mRNAs are poly adenylated. Approximately 200 bp of histone gene 5' flanking sequences are sufficient to confer correct cell cycle regulation on a linked reporter gene indicating that the clustering of histone genes is not important for cell cycle regulation (Schumperli 1988). Possibly there are additional regulatory elements controlling the whole cluster of histone genes, similar to that found in the β-globin cluster (Talbot et al 1989), which have not yet been identified.

The number of examples of clustered genes in the mammalian genome which are not related by DNA or protein sequence is rising rapidly and includes genes which overlap. Where the two genes are transcribed from a bidirectional promoter they probably share transcription elements but in no other case has the clustering or overlapping of the genes been shown to be functionally important. The histone genes are the only example of clustered unrelated genes where the expression of the genes is clearly related. However, the clustering has not been shown to be important to the expression of the histone genes and it is not clear what mechanism has maintained the close proximity of these genes.

Clustered genes in drosophila

Although drosophila has a significantly different genome organization to mammals, the control of expression of genes is very highly conserved and if clustering had a functional significance this might be reflected in the drosophila genome.

The *dunce* gene in Drosophila spans 100 kb of genomic DNA and at least two, and possibly three genes are located inside a large intron of 79 kb (Chen et al 1987). The *dunce* gene encodes the enzyme cAMP phosphodiesterase and is involved in memory. The gene *Sgs-4* is transcribed in the same direction as dunce and is located completely within the second exon. *Sgs-4* is a component of the glue produced in the larval salivary glands and is expressed primarily in the glands of the late third instar larvae. The transcriptional unit for *Sgs-4* has been completely defined and all the cis-acting elements necessary for normal spatial and temporal specificity have been defined between 840 bp 5' and 130 bp 3' of the transcription unit. Upstream from *Sgs-4*, but still within the 79 kb second intron of *dunce* is a second characterized gene called *Pig-1*. *Pig-1* is transcribed in the opposite direction to *Sgs-4* and *dunce*, and is expressed primarily in larval salivary glands, with maximum expression in the second instar. A third gene called *sam* has also been mapped to the 79 kb intron but the transcription unit of this gene has not been defined. The expression of the three characterized genes is completely independent indicating that promoter interference and antisense inhibition are not significant mechanisms in the control of expression of these genes. It is interesting to note that neither of the internal genes contains introns.

The *Gart* gene in Drosophila contains a pupal cuticle protein gene in its first intron (Henikoff et al 1986). *Gart* encodes three purine pathway enzymatic activities and is expressed throughout development. The pupal cuticle gene is only expressed in the epidermis during the prepupal stage

and is not functionally related to the *Gart* gene. The pupal cuticle gene itself contains an intron. There is no evidence of interaction between the regulatory or transcriptional apparatus of the two genes.

The *z600-gonadal-Eip28/29* gene cluster in Drosophila contains three tandemly arranged overlapping genes (Schulz & Butler 1989). The 3' end of the first gene, *z600* overlaps the 5' end of the second gene *gonadal* (*gdl*) by 91 bp, such that the *z600* poly(A) addition site is in the first exon of *gdl*. The 3' end of the two longer *gdl* transcripts overlaps the 5' end of the third gene, *Eip28/29*, by 62 bp. The *z600* gene is expressed abundantly during early embryogenesis where it is localized to dorsal and posterior regions at cellular blastoderm, and then again in adults at a much lower level. The *gdl* gene specifies four different mRNAs. Two mRNAs of 1300 and 1000 bp in length are expressed solely in embryos and adult female ovaries, whereas mRNAs of 1500 and 1200 bp in length are expressed during late larval development and adult male testes. *Eip28/29* is expressed throughout embryonic, larval and adult development. These three tandemly overlapping genes are regulated in quite distinct ways throughout development, and there is no evidence that the clustering is important for their regulation. However, these genes have not been individually introduced into other positions in the genome to determine whether they are correctly regulated in the absence of the others.

The dopa decarboxylase gene in Drosophila overlaps with a convergently transcribed gene of unknown function by 88 bp at the 3' end (Spencer et al 1986). The two genes are never expressed at high levels in the same tissues at the same time and the levels of the two genes are not inversely related. Dopa decarboxylase is maximally expressed in 18-hr embryos, at each larval moult, at puparition and at adult eclosion. The 3' overlapping transcript, however, is abundant only in adult stages and in 1-4-hr embryos.

No functional significance has been found for the overlapping arrangements of the genes described above and the genes are independently regulated. In some cases if you move individual genes elsewhere they are correctly expressed in the absence of the other gene.

Domains containing coregulated genes

There are several regions of the mammalian genome where a number of genes, unrelated by sequence homology, have been identified. However it is still difficult to evaluate whether domains of the genome are co-regulated in some way because no large portion of the mammalian genome has been mapped for all the genes it contains. The drosophila genome is much smaller than the mammalian genome and it has been possible to identify all the genes in defined regions.

A 315 kb segment of the drosophila chromosome, including the *rosy* and *Ace* loci has been cloned and analysed for genetic activity (Bossy et al 1984). Twenty three recombinant phages covering the entire region were hybridized separately to Northern blots of total RNA from salivary glands, fat bodies, whole larvae and poly(A)+ RNA from heads. A total of 39 different RNA species were detected, distributed unevenly over the 315 kb. The first 63 kb contained 18 transcripts, but the next 156 kb contained only seven transcripts. Overall there is no evidence of neighbouring genes being expressed in the same tissues. This study indicates that the very variable distribution of genes found by using CpG-rich islands to detect genes in mammalian DNA, is also found in drosophila. It also suggests that domains of coregulated genes are not a general feature of genome arrangement.

In mammalian genomes "domains" have been postulated to correspond to the loops seen in partially disrupted chromosomes (Gasser & Laemmli 1987). In this theory, the ends of each domain or loop are anchored to the nuclear scaffold and the topological conformation of the loop of DNA is

regulated. "Domains" have also been postulated to correspond to the G banding seen on Giemsa staining of metaphase chromosomes. It is possible that there is a long range ordering within the euchromatin and that this is visualized by banding techniques.

A "two genome" model has been proposed which links the high resolution (2000 kb) Giemsa banding (G bands) to two types of domains (Goldman et al 1984). The dark G bands are thought to correspond to A+T rich regions which contain tissue specific genes and are late replicating (Holmquist 1987). The light G bands are G+C rich regions which contain constitutively expressed genes and are early replicating (Holmquist 1987). An example of dark G banding DNA would be the 10 Mb of DNA around the Duchenne Muscular Dystrophy gene (Burmeister 1988) and an example of light G banding DNA would be the CpG-rich region located at 4p15-4pter (Pohl et al 1988).

The distribution of genes in the mammalian genome does seem to be very uneven and there may well be a correlation between G bands and the distribution of housekeeping as opposed to differentially expressed genes. However, the banding is likely to be due to gross average differences rather than a functional sorting into differentially controlled domains, in the same way that Drosophila genes are found very unevenly distributed, but the differential expression of neighbouring genes does not appear to be related. Sorting into differentially expressed domains has not been observed at the DNA level where individual neighbouring genes are not coregulated and the genome is remarkably resistant to translocations and rearrangements which would interrupt such domains.

The α- and β-globin gene clusters are possible examples of coregulated domains. The human β-globin cluster contains 5 related genes called ε, Gγ, Aγ, δ and β spread over about 60 kb. During development the genes are successively expressed starting at the left of the cluster with ε-globin which

is expressed in the yolk sac and ending with δ- and β-globin which are expressed in adult bone marrow (Maniatis et al 1980). This suggests that the order of the genes is related to their successive activation and the switch between the chicken ε- and β-globin genes is a mechanism which could cause sequential expression of the genes (see section "Clusters of repeated genes"). However, individual human β-globin genes with several kb of flanking DNA are correctly expressed in transgenic mice during development (Kollias et al 1986, 1987). Even though the transgenes are correctly differentially expressed during development, the expression of the transgenes is at a lower level than the endogenous genes and they are susceptible to position effects. DNase hypersensitive sites specific for erythroid cells are located on either side of the β-globin cluster and represent additional regulatory regions (Grosveld et al 1987, van Assendelft 1989). A 6.5 kb fragment of DNA encompassing the hypersensitive sites located 5' to the ε-globin gene confers position independent expression on a linked β-globin gene (Talbot et al 1989). This 6.5 kb of DNA is normally located about 50 kb from the β-globin gene and on the far side of four other genes and it might regulate the entire cluster from its normal location.

Distribution of coregulated genes.

Clusters of duplicated genes are often coregulated or regulated only slightly differently and this is probably due to the duplication of the control elements rather than the existence of coregulated domains. There are plenty of examples of duplicated, neighbouring genes whose control is completely different, and individual genes from a cluster can be transfected to different positions in the genome and are correctly expressed. Conversely, co-expressed groups of genes which have not arisen recently by duplication, are not clustered in the genome. For example GM-CSF, interleukin 2 and interferon-γ are transcribed concomitantly following T cell activation but

are located on human chromosomes 5, 4 and 12 (Barlow et al 1987). The four mammalian ribosomal protein genes, L32, L30, S16 and L7a map to mouse chromosomes 6, 15, 7 and 2 respectively (Wiedemann et al 1987, section 7.4 this thesis). The enzymes galactose-4-epimerase, galactokinase and galactose-1-phosphate uridyltransferase which are all involved in galactose metabolism are located on human chromosomes 1p, 17q and 9p respectively (John & Miklos 1988 p197). There are numerous other examples of muscle, liver and kidney specific genes which are at widely spread out locations on the genome.

Conclusion

In general mammalian genes are independently regulated by transacting protein factors (Dynan & Tjian 1985). Coregulated genes which have arisen by gene duplication are often located in clusters and these genes are largely independently regulated. Whereas coregulated genes which have not arisen by duplication, are widely dispersed in the genome. The enormous influx of repetitive elements has not upset regulation by any of the postulated mechanisms of interference of transcription machinery or antisense inhibition.

There are four mechanisms unique to eukaryotes which conserve regulatory DNA and rely on the *cis* location of genes. These are co-expression of the rRNA genes by rRNA polymerase I, differential splicing of genes to give more than one protein product, rearrangement of the immunoglobulin genes and inactivation of the X chromosome in female mammals.

In addition there are individual cases where the juxtaposition of genes is used for regulation of gene expression. Bidirectional promoters located in CpG-rich islands are used to drive transcription of two genes, nearby genes can share enhancer elements, promoter occlusion is used to switch between

two promoters driving transcription of one gene and dominant activating sequences may activate clusters of genes. However there are no examples where a piece of DNA is transcribed from both strands and this causes antisense inhibition.

Possibly the flexibility of the eukaryotic genome which allows shuffling of exons (Gilbert 1978) and the duplication of genes and their subsequent recruitment to new functions, precludes the widespread use of *cis* interactions between genes. This flexibility has also allowed the massive influx of highly repetitive elements and the shuffling of DNA which has occurred between mammalian species (Nadeau & Taylor 1984), neither of which appear to be beneficial. Although *cis* interactions are not widely used, and the genome is resistant to such effects caused by dispersed repetitive elements, the molecular mechanisms do work in mammalian cells and are used in individual cases to regulate gene expression. It is likely that as the regulation of more closely juxtaposed genes is analysed in detail, more examples of the use of these mechanisms in gene control will be found.

Yeast has a very small genome, presumed to be selected to be small so that yeast can replicate rapidly. However, even though *cis* interactions are economical on regulatory DNA, there is no more evidence of *cis* interactions between genes in yeast than in mammals.

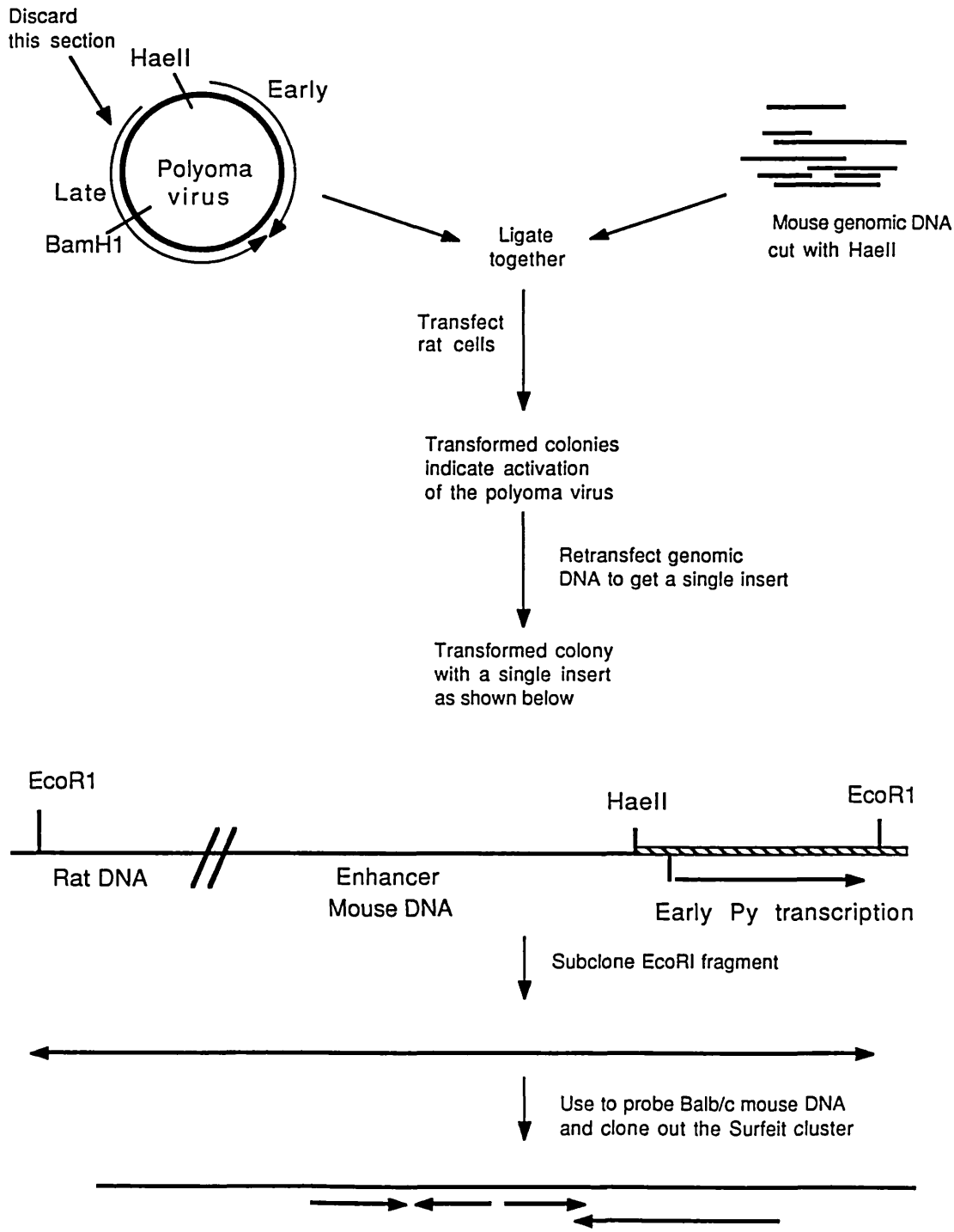
1.3 Cloning the Mouse Surfeit Gene Cluster

The work presented in this thesis arises out of an experiment which was designed to clone fragments of mouse genomic DNA which act as an enhancer of transcription, without any prior knowledge of the genes which such an enhancer might be associated with. This allows one to clone functionally important DNA from the mammalian genome without restricting oneself to highly expressed genes or genes of defined function which can be cloned out by other means. At the same time this technique ignores the vast amount of "junk" DNA in the mammalian genome, including satellite repeat DNA, intron sequences, pseudogenes etc. Having cloned an enhancer sequence it is then of interest to know how it relates to the expression of nearby genes and whether these genes have the same structural features as mammalian genes cloned out by other means.

The experiment to clone enhancers by expression selection is fully described in Fried et al (1983) and the strategy is summarized in figure 1.1. The Polyoma virus early region causes transformation in rat cells, and the enhancer required for expression of the early genes is located between the HaeII and BamHI sites upstream of the promoter region. If this enhancer fragment is deleted the early genes of the resulting viral molecule are poorly expressed upon transfection into rodent cells even though 50 bp including the early transcription unit TATA box is retained. Thus the number of transformed colonies observed in such a transfection is about 0.1% of that observed with the intact viral DNA. To select for DNA with enhancer activity, mouse DNA was digested with HaeII and ligated in place of the viral enhancer fragment onto the enhancerless Polyoma DNA vector. This DNA, when transfected into Rat cells can cause transformation when the ligated mouse DNA activates the expression of the polyoma virus early genes. The DNA was transfected into Rat cells by $CaPO_4$ precipitation which often causes multiple inserts of transfected DNA into the genome. To obtain a single insert, the

genomic DNA from a primary focus was retransfected into rat cells and a secondary transformed colony picked. One particular mouse enhancer initially called H2, which has since been renamed MES 1 (Mouse Expression Sequence 1) was cloned in this way.

Figure 1.1 Isolation of the MES-1 Expression Sequence

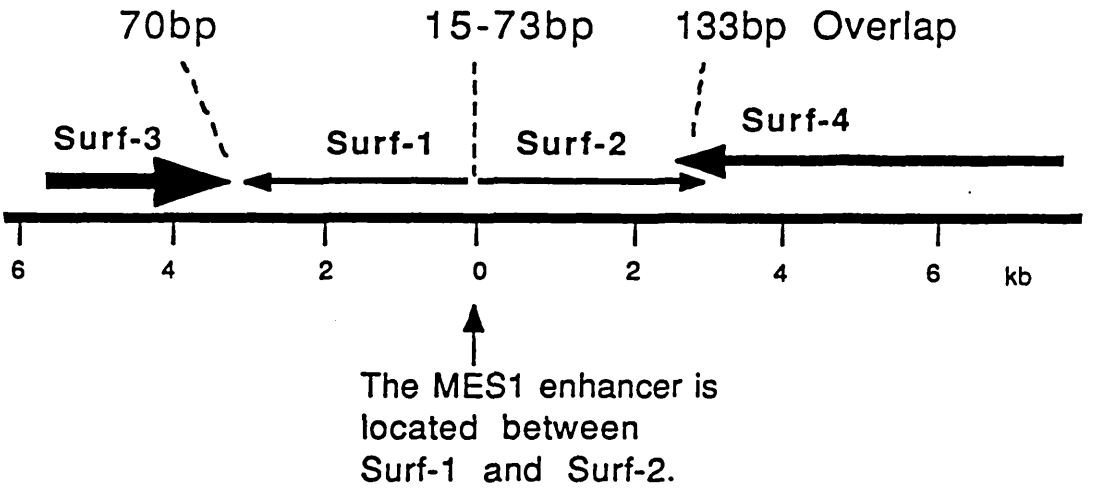


1.4 Organization of the Mouse Surfeit Gene Cluster

Having cloned the MES 1 element, it was of interest to see how it was associated with genes in its normal genomic location. Figure 1.2 summarizes the layout of four genes which form a cluster around the MES 1 element, the whole region being called the Surfeit cluster, or locus, and the genes being called Surf-1 to -4. MES 1 is located between the 5' ends of Surf-1 and Surf-2 which are described in Williams and Fried (1986a). The heterogeneous 5' ends of these genes are separated by just 15-73 bp and are located in a CpG-rich island which does not contain any TATA boxes. Surf-4 is transcribed towards Surf-2 and the 3' ends of these genes overlap by 133 bp (Williams and Fried 1986b). Surf-3 is transcribed towards Surf-1 with its 3' end 73 bp from the 3' end of Surf-1 (Williams 1986). Surfs -1 and -2 are expressed at a low level, about 0.01% of total poly(A)+ RNA in a variety of cell lines (Williams et al 1988) and they can both express more than one mRNA as a result of different splicing. The most abundant mRNA in the case of Surf-1 encodes a putative protein of 257 amino acids, and in the case of Surf-2 a protein of 306 amino acids. The function of the Surf-1 and -2 putative proteins is at present unknown. Surf-3, which is discussed in detail in chapter 2, is a member of a multigene family (Huxley et al 1988), and specifies a very abundant mRNA accounting for about 0.1% of poly(A)+ RNA in a variety of cell lines (Williams et al 1988). Surf-4 is moderately abundant, accounting for about 0.05% of total poly(A)+ RNA in a variety of different cell lines, but it is down regulated in undifferentiated teratocarcinoma cell lines (Williams & Fried 1986b). Surf-4 is described in detail in chapter 3.

Figure 1.2 Organization of the Surfeit Cluster.

The solid line indicates the cloned mouse genomic DNA with distances shown in kb. The arrows above show the extent and direction of each Surfeit gene without showing exon intron structure.



1.5 Significance of the Clustering of the Surfeit Genes

The Surfeit genes lie in a very tight cluster of alternating transcription units with no more than 73 bp separating any two genes. This is in marked contrast to all previously cloned mammalian genes (see section 1.2) and suggests that the organization of the genes may be related to the control of their expression. There are several different molecular mechanisms which could be involved in control of gene expression and which rely on the *cis* location of the genes (see section 1.1). The bidirectional promoter may direct co-ordinate or alternate expression of Surf-1 and Surf-2. Surf-1 and Surf-2 may share regulatory elements, or the regulatory elements of one gene may be located in the other gene. The 3' overlap between Surf-2 and Surf-4 may allow antisense control. The 3' ends of Surf-3 and Surf-1 do not overlap but RNA polymerase II does not stop at the poly(A) addition site so expression of Surf-3 may interfere with expression of Surf-1. Surf-3 is transcribed in the same direction as Surf-2, and Surf-4 is transcribed in the same direction as Surf-1, allowing promoter occlusion to occur between these pairs of genes.

The aim of the work presented in this thesis was to identify the other Surfeit genes and to understand how they are regulated and why they should be arranged as they are. This thesis describes the organization and sequence of Surf-3 and Surf-4 and the identification of the other members of the Surf-3 multigene family as processed pseudogenes, the identification of two new Surfeit genes, the location of four CpG-rich islands within the Surfeit cluster and the location of two flanking CpG-rich islands. The cell cycle regulation of the Surfeit gene expression has been investigated and the Surfeit gene cluster has been mapped in the mouse.

Chapter 2: Structural Analysis of Surf-3

2.1 Introduction

The Surf-3 transcription unit lies to the left of the Surfeit cluster being transcribed towards Surf-1 (Fig. 2.1). Hybridization of a cDNA of Surf-3 to Northern blots shows that Surf-3 is homologous to a very abundant 1.0 kb mRNA (Fig. 2.1a) which accounts for about 0.1% of the total poly(A)+ RNA in a variety of cell lines (Williams 1986). Although the Surf-3 cDNA hybridizes to a single size of mouse mRNA, a Southern blot of restricted mouse cellular DNA probed with this cDNA shows numerous bands homologous to Surf-3 (Fig. 2.1b). The bands on a southern blot hybridizing to different fragments of the cDNA were very similar showing that there are multiple sequences in the genome highly homologous to the whole Surf-3 gene and that the detection of multiple bands is not due to a repetitive element in the cDNA (Williams 1986). It was further shown that Surf-3 is also a member of a similar multigene family in human, monkey and rat DNA (Williams 1986).

As Surf-3 is a member of a multigene family it could be that the mRNA detected by Northern blotting is actually derived from another region of the genome rather than the Surfeit cluster. To address this, the mouse 19 kb EcoRI fragment containing Surf-3, attached to the SV40 origin of replication, was transfected into COS monkey cells. The mouse Surf-3 mRNA was found to be highly abundant in the transfected monkey cells showing that Surf-3 is an expressed gene, however, it does not rule out that the other copies of the Surf-3 multigene family are also expressed (Williams 1986). A number of mouse Surf-3 cDNAs, including C2C, C3E and C4G have been isolated from RNA extracted from the mouse Surf-3 transfected monkey cells, and a Surf-3 cDNA called MY3 has been isolated from Balb/C TS-A-3T3 poly(A)+ RNA. The cDNAs MY3 and C2C have largely been sequenced (Williams 1986).

Any clue as to the identity of one of the Surfeit genes will aid greatly in our understanding of the regulation of these genes. Surf-3 is a very highly expressed gene and might therefore be more easily identified than the other Surfeit genes. Sequencing the cDNAs derived from mouse cellular RNA would also tell us whether the mouse mRNA observed on Northern blots is predominantly derived from Surf-3 or from another, somewhat diverged, member of the multigene family.

Determining the layout of the Surf-3 gene and the sequence of the 5' flanking region gives information about the layout of the entire Surfeit locus and may tell us about regulation of Surf-3 transcription, which we can relate to the whole locus. Some of the other members of the Surf-3 multigene family were cloned and analyzed to determine their relationship to Surf-3.

2.2 Results: Sequence and Organization of Surf-3

The sequencing of Surf-3 cDNA MY3 and a mouse cDNA called C2C derived from poly(A)+ RNA of monkey COS cells transfected with the 19 kb EcoRI fragment, initiated by T. Williams, was completed (see appendix 1). The fragments of the cloned 19 kb EcoRI fragment homologous to the exons of Surf-3 were identified by Southern blotting and these were in turn sequenced (see appendix 1). The overall sequencing determined that MY3 is homologous to sequences in the 19 kb EcoRI genomic DNA fragment except for one base pair, and spread over 8 exons covering 3 kb of genomic DNA (Fig. 2.1). The cDNA C2C is exactly homologous to these 8 genomic exons, differing from MY3 at one nucleotide (see above). This single nucleotide which is different in MY3 from the cDNA C2C and from the genomic sequence could be due to a mutation caused during reverse transcription and generation of the cDNA MY3, or due to a difference between the Balb/C TS-A-3T3 cell line and the Balb/C A31 cell line.

The sequencing shows that Surf-3 (or an almost completely homologous gene), is highly expressed in mouse, and the mRNA observed on Northern blots is not derived from a diverged member of the multigene family. Furthermore, C2C is a mouse rather than monkey cDNA confirming that Surf-3 is transcribed from the mouse 19 kb EcoRI genomic DNA fragment on transfection into monkey COS cells. Surf-3 contains a single long open reading frame which encodes a polypeptide of 266 amino acids (Fig. 2.2). The Surf-3 protein is very basic (24% Arg + Lys) but has no other obvious features such as transmembrane regions or ATP binding sites. Surf-3 also had no particular homology to any DNA sequences in the EMBL and GENBANK databases, or to any proteins in the PIR databases.

The Surf-3 cDNAs all end within a few bases of the G at -7 in figure 2.2 indicating that this is either the 5' end of the gene and the 200 bp difference in size between the cDNA and the mRNA is due to the poly(A) tail, or there is a very strong block to reverse transcriptase at this point, and the 5' end of the mRNA has not been cloned. To determine the 5' end of the Surf-3 mRNA accurately, primer extension and RNase protection experiments were carried out. For primer extension, a 30 bp oligonucleotide, homologous to the 5' end of the cDNAs was utilized in the hope that this would evade any secondary structure present in the mRNA which could be causing premature termination. The results of primer extension analysis using this end labelled oligo with Balb/C TS-A-3T3 poly (A)+ RNA and Yeast tRNA are presented in figure 2.3. The results indicate that there are two major start sites of Surf-3 transcription in the prominent run of pyrimidines just upstream to the 5' ends of the cDNAs.

For RNase protection experiments a 200 bp FnudII fragment spanning the 5' most exon present in the cDNAs was cloned into the vector pSP64 behind the SP6 promoter, such that the SP6 generated RNA would be complementary to the Surf-3 transcript. The major fragments generated by

protection of this probe by mouse poly(A)+ RNA from RNase digestion were 25 and 29 bp in length. Taking into account the position of the first splice site, this corresponds to the same two major start sites observed with primer extension (Fig. 2.4). The RNase protection results are less precise than the primer extension because of the presence of background fragments only slightly shorter than the protected fragments, even with the tRNA control.

The finding that the major sites detected by primer extension and RNase protection are the same, strongly indicates that these sites correspond to start sites of transcription rather than splice sites or sites of secondary structure. The sequence of the promoter is discussed in greater detail in section 2.4.

Figure 2.1 Northern and Southern blot analysis of Surf-3.

At the top is a map of the Surfeit cluster with the Surf-3 gene enlarged below. The continuous horizontal line represents the genomic DNA with distances indicated in kb. The four characterized Surfeit genes are shown by arrows, with the direction and extent of the arrows indicating the direction and extent of transcription. Below is show an enlargement of the Surf-3 transcription unit. The exons are indicated by solid boxes and the introns by the thinner line.

a) Northern blot probed with cDNA MY3.

The first track contains 20 µg of Poly(A)- RNA, and the second track 1 µg of Poly(A)+ RNA from mouse Balb/C TS-A-3T3 cells. The size of the mRNA detected was calculated from the positions of the 18S and 28S rRNAs.

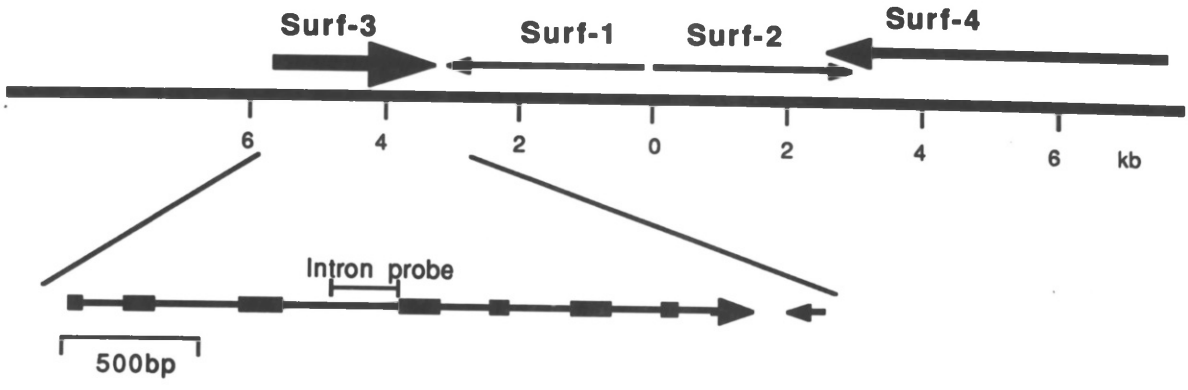
b) Southern blot of mouse DNA probed with cDNA MY3.

Each lane contains 10 µg of mouse Balb/C A31 DNA cut with EcoRI or BamHI.

c) Southern blot of mouse DNA probed with an intron of Surf-3.

Each lane contains 10 µg of DNA from Balb/C A31 cells cut with EcoRI, HindIII or BamHI and the blot is probed with the intron probe indicated on the map above above. This probe is a BamHI to HindIII fragment and is shown in more detail in appendix 1a.

The size markers for the Southern blots were HindIII fragments of Lambda phage DNA.


Surf-3
Surf-1
Surf-2
Surf-4
6
4
2
0
2
4
6
kb
Intron probe
500bp

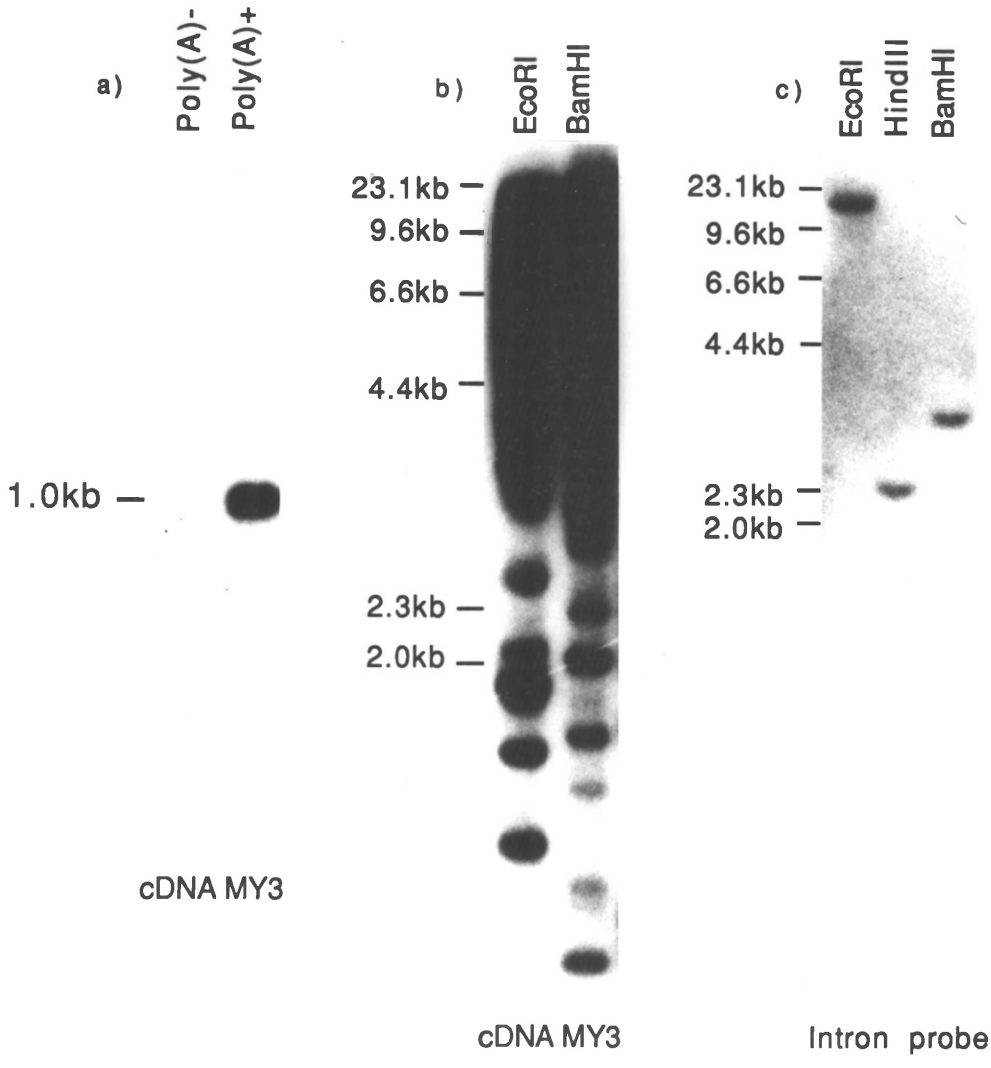

a)
Poly(A)-
Poly(A)+
1.0kb
cDNA MY3
b)
EcoRI
BamHI
23.1kb
9.6kb
6.6kb
4.4kb
2.3kb
2.0kb
cDNA MY3
c)
EcoRI
HindIII
BamHI
23.1kb
9.6kb
6.6kb
4.4kb
2.3kb
2.0kb
Intron probe

Figure 2.2 DNA sequence and protein encoded by Surf-3.

The sequence of the Surf-3 cDNA is given in the top line with the amino acids coded by the long open reading frame given below. The T at position 544 is a T in the genomic DNA and cDNA C2C, but a C in cDNA MY3. The non-consensus poly(A) addition site, AATATA is underlined, the intron positions are marked and the * marks the 5' end of cDNA MY3. Numbering is from the ATG initiation codon.

```
                    *       1  intron 1                          33
TTCTTTCTCCAGCAGCCGAGCAAGATGCCCAAAGGGAAGAAGGCCAAGGGGAAGAAG
                          METProLysGlyLysLysAlaLysGlyLysLys
                                                               90
GTGGCCCCGGCCCCCGCCGTCGTCAAGAAGCAGGAGGCCAAGAAGGTGGTCAACCCT
ValAlaProAlaProAlaValValLysLysGlnGluAlaLysLysValValAsnPro
                                          intron 2             147
CTGTTCGAGAAGAGGCCCAAGAACTTCGGCATTGGACAGGACATCCAGCCCAAAAGA
LeuPheGluLysArgProLysAsnPheGlyIleGlyGlnAspIleGlnProLysArg
                                                               204
GATTTAACGCGCTTCGTCAAATGGCCCCGCTACATCAGACTGCAGCGGCAAAGAGCT
AspLeuThrArgPheValLysTrpProArgTyrIleArgLeuGlnArgGlnArgAla
                                                               261
ATCCTCTACAAGCGGCTCAAAGTCCCTCCTGCCATTAACCAGTTCACCCAGGCCCTG
IleLeuTyrLysArgLeuLysValProProAlaIleAsnGlnPheThrGlnAlaLeu
             intron 3                                          318
GACAGGCAGACAGCTACTCAGCTGCTTAAGCTTGCCCACAAGTACAGGCCAGAGACA
AspArgGlnThrAlaThrGlnLeuLeuLysLeuAlaHisLysTyrArgProGluThr
                                                               375
AAGCAAGAGAAGAAGCAAAGGCTACTGGCCCGTGCTGAGAAGAAAGCTGCTGGCAAA
LysGlnGluLysLysGlnArgLeuLeuAlaArgAlaGluLysLysAlaAlaGlyLys
                                              intron 4         432
GGCGACGTCCCAACTAAGAGACCACCTGTCCTCCGAGCAGGAGTCAATACAGTCACC
GlyAspValProThrLysArgProProValLeuArgAlaGlyValAsnThrValThr
                                                               489
ACCTTGGTGGAGAACAAGAAGGCTCAGCTGGTGGTGATTGCCCATGACGTAGACCCC
ThrLeuValGluAsnLysLysAlaGlnLeuValValIleAlaHisAspValAspPro
      intron 5                                                 546
ATTGAGCTGGTGGTTTTCCTACCTGCTCTGTGTCGAAAGATGGGGGTGCCCTACTGC
IleGluLeuValValPheLeuProAlaLeuCysArgLysMetGlyValProTyrCys
                                                               603
ATCATCAAGGGAAAGGCCAGGCTGGGGCACCTGGTCCACAGGAAGACATGCACCACC
IleIleLysGlyLysAlaArgLeuGlyHisLeuValHisArgLysThrCysThrThr
                              intron 6                         660
GTTGCCTTCACACAGGTTAACTCGGAAGACAAGGGTGCTCTGGCTAAGCTGGTGGAA
ValAlaPheThrGlnValAsnSerGluAspLysGlyAlaLeuAlaLysLeuValGlu
                                                  intron 7     717
GCTATTAGGACCAATTATAATGACAGATATGACGAGATCCGTCGCCACTGGGGAGGC
AlaIleArgThrAsnTyrAsnAspArgTyrAspGluIleArgArgHisTrpGlyGly
                                                               774
AACGTCCTGGGTCCTAAGTCTGTGGCTCGAATTGCCAAGCTGGAAAAAGCAAAGGCT
AsnValLeuGlyProLysSerValAlaArgIleAlaLysLeuGluLysAlaLysAla
                                                               831
AAAGAACTCGCCACTAAATTGGGTTAAATGTACACTAAATTTTCTGTACCTAAATAT
LysGluLeuAlaThrLysLeuGly
          844
AATTACAAAATTA
```

Figure 2.3 Primer Extension analysis of the 5' end of Surf-3.

The sequence at the 5' end of the Surf-3 gene is indicated above. The sequence runs from upstream of Surf-3, through the first exon and into the second exon with the position of the first intron indicated. The 30 bp oligonucleotide used for primer extension is underlined, and the start sites which correspond to the extended products of 52 and 48 bp are indicated by *

Below is shown a primer extension experiment. The first two lanes show primer extension carried out with 1 μg of Poly(A)+ RNA from Balb/C TS-A-3T3 cells and 100,000 cpm of ^{32}P end labelled oligonucleotide. The third and fourth lanes are control primer extensions using tRNA instead of poly(A)+ RNA to show that the extension products are specific for the mouse RNA. The fifth lane is empty, and the sixth lane shows the end labelled 30 bp oligonucleotide without added reverse transcriptase. For this experiment, the labelled oligonucleotide was separated from unincorporated nucleotides by passage through a G50 spun column. There is an appreciable contamination with non full-length oligonucleotides and a 55-56 bp oligonucleotide which is present in all the lanes. The right hand four columns show a sequencing ladder generated with single stranded M13 DNA and the standard 15 bp M13 primer.

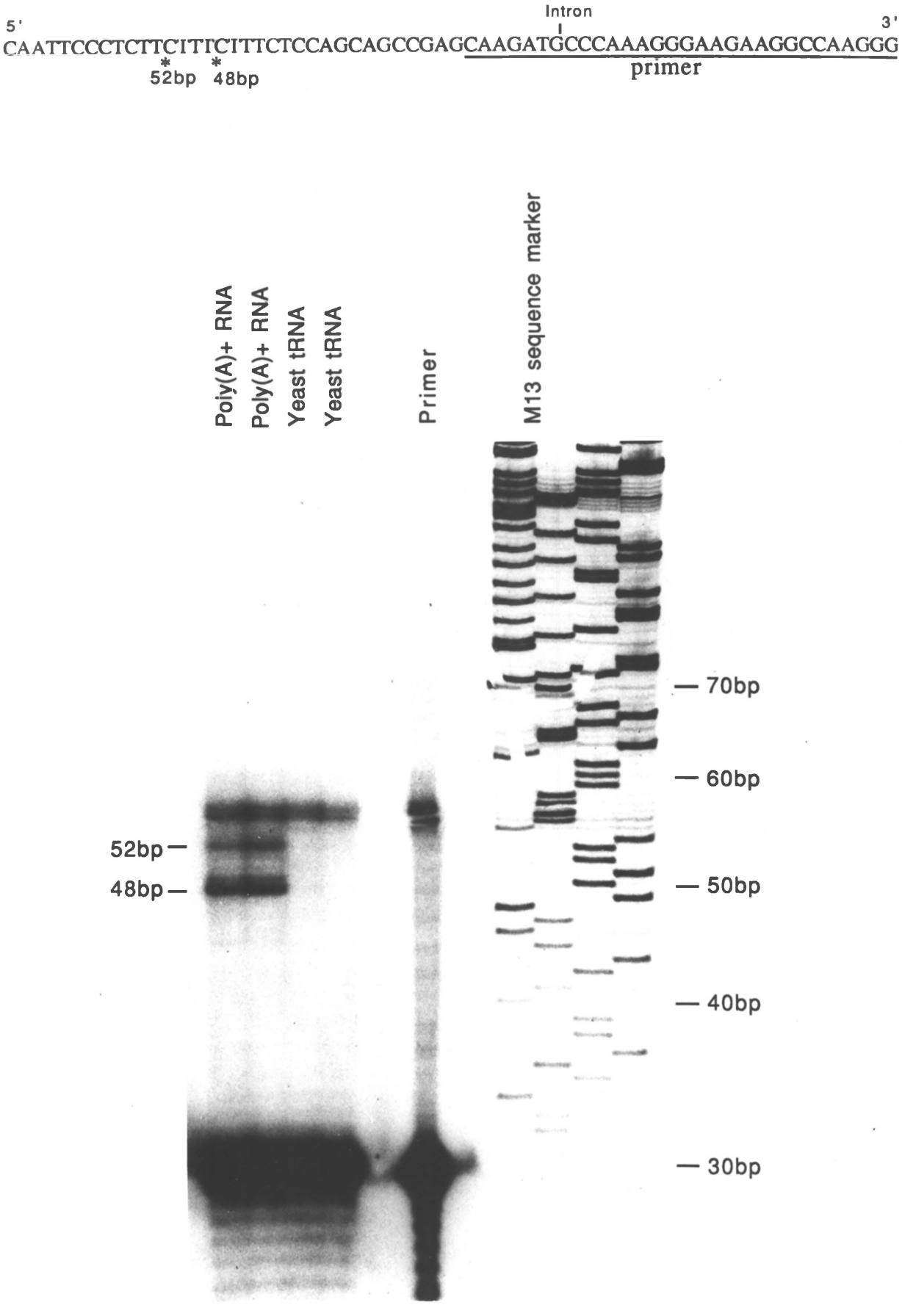
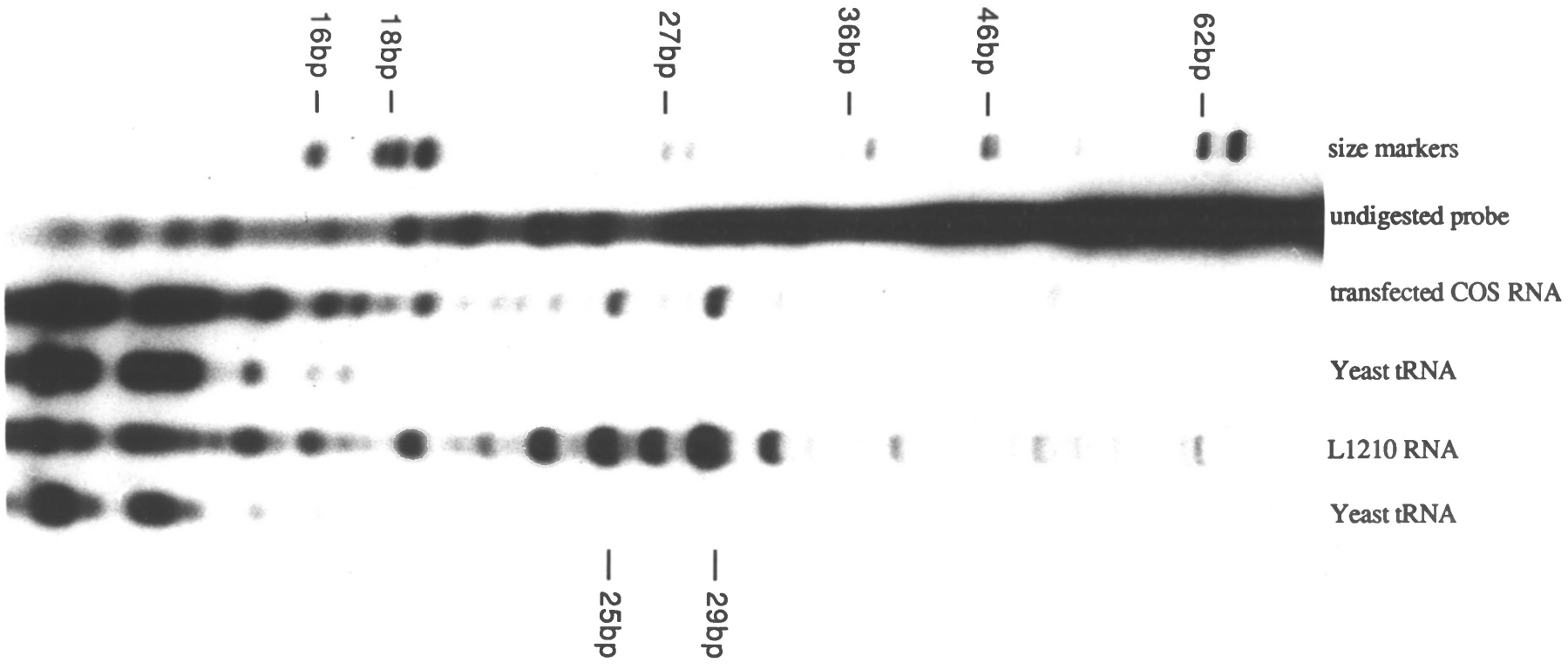

Intron
5'
3'
CAATTCCCTCTTCITTCITTCTCCAGCAGCCGAGCAAGATGCCCAAAGGGAAGAAGGCCAAGGG
primer
52bp
48bp
Poiy(A)+ RNA
Poly(A)+ RNA
Yeast tRNA
Yeast tRNA
Primer
M13 sequence marker
52bp
48bp
70bp
60bp
50bp
40bp
30bp

Figure 2.4 RNase protection analysis of the 5' end of Surf-3.

To the right is shown the genomic sequence spanning the first exon of Surf-3. The first splice site and the positions of the start sites corresponding to the protected fragments of 25 and 29bp are indicated.

On the left is shown an RNase protection experiment. A 200 bp FnudII genomic fragment spanning the first exon was cloned in pSP64 in the orientation such that SP6 polymerase generated RNA transcripts are complementary to the Surf-3 mRNA. Single stranded RNA transcripts continuously labelled with ^{32}P UTP were purified from a denaturing polyacrylamide gel. For each RNase protection 100,000 cpm of SP6 generated RNA was used. The lefthand lane shows size markers made from end labelled DdeI fragments of a plasmid containing a fragment of polyoma virus. The sizes of the fragments are shown on the left, but these will not correspond exactly in size with co-migrating RNase protected fragments because single stranded DNA and single stranded RNA migrate with slightly different mobilities on a denaturing polyacrylamide gel. The second lane contains undigested labelled single stranded RNA probe, this probe has undergone considerable degradation due to the high specificity of labelling used and the large amount of probe used to detect the short protected fragments expected. The other lanes contain RNase protected fragments remaining after annealing of the single stranded RNA probe to 1 µg of poly(A)+ RNA from monkey COS cells transfected with the mouse 19 kb EcoRI fragment containing Surf-3, 1 µg of poly(A)+ RNA from mouse DBA L1210 cells or 40 µg of yeast tRNA.



5' 1st exon Surf-3 3'

CACAATTCCCTCTTCTTTCTTTCTCCAGCAGCCGAGCAAGATGGTGAGTGTCTCT

29bp 25bp

splice site

2.3 Results: Cloning and Sequencing Surf-3 Pseudogenes

Surf-3 is a member of a multigene family (Fig. 2.1b). When a probe from the third intron of Surf-3 is used to probe a Southern blot of restricted mouse genomic DNA however, it is found to hybridize to unique fragments (Fig. 2.1c), indicating that the other members of the family are not recently formed duplications. To determine the nature of the other genomic fragments homologous to Surf-3, a genomic library of EcoRI fragments (not size fractionated) was constructed in lambda GT10. Lambda GT10 is an insertion vector and will clone fragments up to 9 kb in size efficiently. As the EcoRI fragment containing Surf-3 is 19 kb in size, this should not be represented in the library. However, as can be seen in figure 2.1b a number of the other members of the multigene family are contained in EcoRI fragments smaller than 9 kb and they should be included in the library. A library of 50,000 recombinant lambda phage plaques was screened with the 5' end of Surf-3 cDNA C4G and 6 positives were isolated, subcloned and analysed further. Each positive clone was mapped with a variety of restriction enzymes and southern blots of fragments were probed with first the 5' end of Surf-3 and then the 3' end to determine the position and extent of the Surf-3 homologous sequences. Maps of the Lambda clones and the extent of homology with Surf-3 are shown in appendix 1c.

When fragments of the positive clones were sequenced it became clear that all the clones investigated contain processed pseudogenes. The sequencing strategy is shown in appendix 1c. Figure 2.5 shows the 5' sequence of clones 3, 4 and 5. A comparison of the sequences with the Surf-3 sequence shows that the first intron is missing in every case and although the transcribed region of Surf-3 is highly conserved in the pseudogenes, homology ceases beyond the 5' end of the gene. Figure 2.6 shows the sequences of the complete pseudogenes in clones 1, 2 and 6. These

pseudogenes have all the hallmarks of having been formed by reverse transcription and subsequent reintegration into the genome. At the 3' end of each pseudogene there is a stretch of As which is thought to be a residue of the poly(A) tract of the mRNA. At the 5' and 3' end of each pseudogene, there is a direct repeat of 13 to 15 bp which is different in each pseudogene. These repeats are thought to be formed during integration of the pseudogene and they define the 5' end of the pseudogene very precisely. In the majority of processed pseudogenes studied, the 5' end of the pseudogene coincides with the CAP site of the mRNA (Weiner et al 1986). Here, the 5' ends of the pseudogenes coincide with the region spanned by the major start sites of transcription of Surf-3 as defined by primer extension and RNase protection (see section 2.2).

All three complete pseudogenes are highly homologous to Surf-3, ranging from 82-95% identity, and if it is assumed that the pseudogenes are under no selection pressure, and that substitutions take place at roughly 5 x 10^{-9} per site per year, then this corresponds to some 10 to 25 million years of divergence (Kimura 1983). All 3 pseudogenes contain stop codons in the reading frame which is used in Surf-3 and would not be able to code a protein of significant size. The first ATG of the pseudogene in clone 6 is mutated to ATA.

Figure 2.5 The 5' ends of three pseudogenes of Surf-3.

The top line gives the sequence of the Surf-3 exons and the Surf-3 upstream genomic sequence. The two *s indicate the two major start sites as defined by S1 mapping and primer extension (section 2.2). The three lines underneath are the sequences of three of the pseudogenes. - indicates identity with Surf-3, ^ indicates a gap inserted in one of the pseudogene sequences to maintain the alignment with Surf-3 and extra nucleotides in the pseudogene are shown displaced one line down. The restriction maps and sequencing strategy of the lambda clones are given in appendix 1c. The top pseudogene in this figure is from lambda clone 3, the middle is clone 4, and the bottom one is clone 5.

```
                      *   *                          intron 1
TCCCCCACAATTCCCTCTTCTTTCTTTCTCCAGCAGCCGAGCAAGATGCCCAAAGGGAAGAAGGCC
TTTTTTTTAAAGGCTTGGGGAGGCGC------C----------------------A---------
TCGTGCAAGAACCCTCTCTCC-G-------------------------------------------
TTTTTAAAAAATTATCAAAT--G--C------C----------------------A-----A---


AAGGGGAAGAAGGTGGCCCCGGCCCCCGCCGTCGTCAAGAAGCAGGAGGCCAAGAAGGTGGTCAAC
--------------------A-TT---A-T---A-------A-----------A-----------T
-----------------------------------A-------------------------------
-----A-------------^------T---AC-A-------AT--T---T---A--A--------T
                                                          A

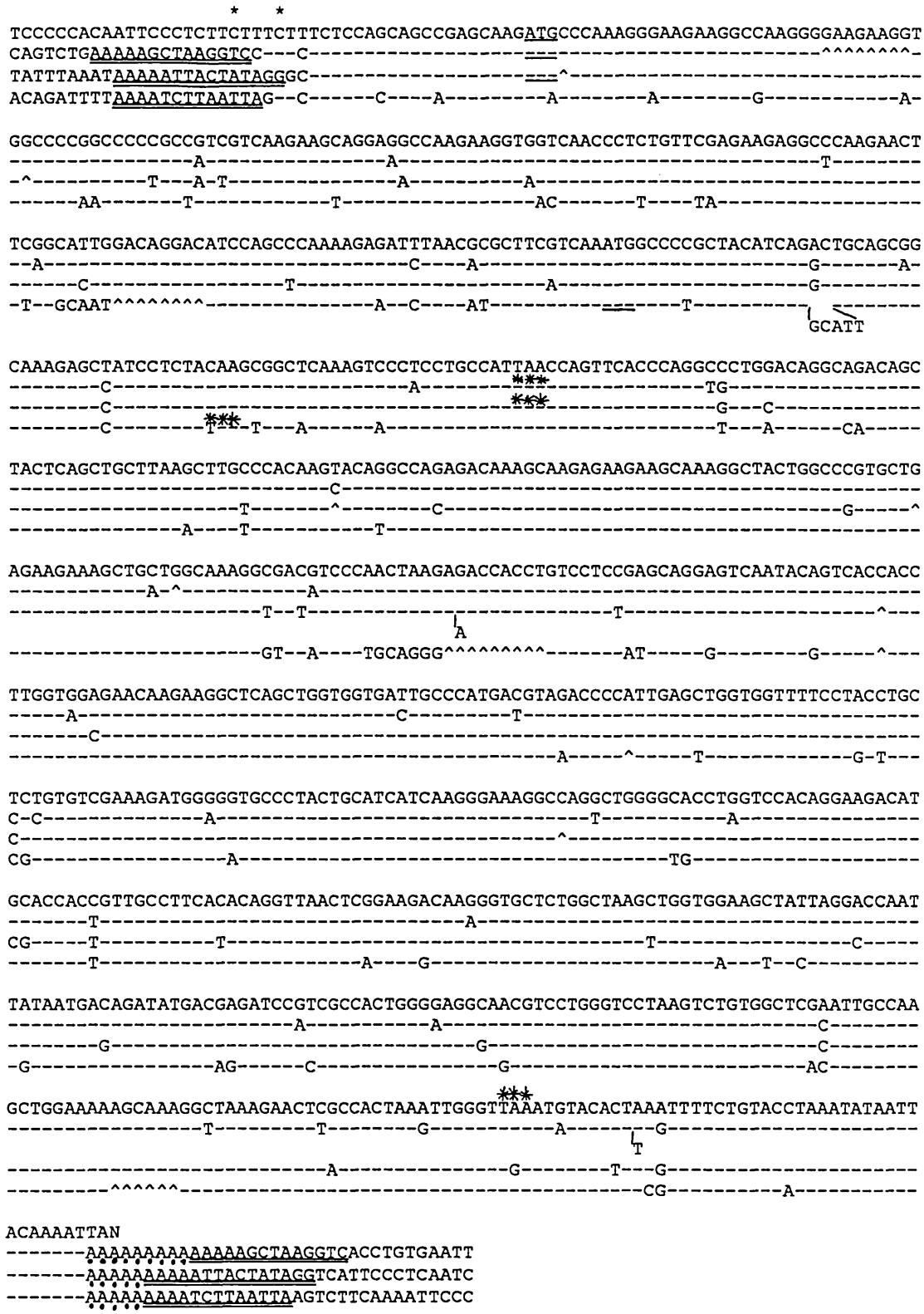

                                          intron 2
CCTCTGTTCGAGAAGAGGCCCAAGAACTTCGGCATTGGACAGGACATCCAGCCCAAAAGAGATTTA

---T----T-----A----------------C-----AGA----------------------C-
------------------------------------------------------------------
---T----T-----A--------------T--------G--------T----------------C-


ACGCGCTTCGTCAAATGGCCCCGCTACATCAGACTGCAGCGG
------------------T-------------G------
```

Figure 2.6 Three completely sequenced pseudogenes of Surf-3.

The top line shows the sequence of the Surf-3 cDNA and genomic sequence extending 5' to the Surf-3 gene. The two *s indicate the two major start sites as defined by S1 mapping and primer extension. The three lines underneath are the sequences of three complete pseudogene. - indicates identity with Surf-3, ^ indicates a gap inserted in one of the pseudogene sequences to maintain alignment with Surf-3 and extra nucleotides in the pseudogene sequences are shown displaced one line down. The direct repeats at the 5' and 3' ends of each pseudogene are underlined twice and the residual poly(A) tract is underlined with dots. For each gene the first inframe ATG is underlined and the first inframe termination codon is overlined with ***. Appendix 1c shows the restriction maps for these three pseudogenes, the top pseudogene in this figure, is from lambda clone 1, the middle from lambda clone 2, and the bottom line is clone 6.



2.4 Discussion: What is the Surf-3 Protein?

Surf-3 is a highly transcribed gene, which is a member of a multigene family. Whereas Surf-3 is an intron containing gene which is highly transcribed when transfected into monkey COS cells, the other members of the family are predominantly, if not all, processed pseudogenes which are probably not transcribed. Even if the pseudogenes were transcribed they would not be able to encode the Surf-3 protein because of base changes, additions and deletions (Fig 2.5 & 2.6). Surf-3 encodes a very basic protein of 266 amino acids. This taken together with its being a member of a large multigene family and its high expression, suggests that two candidates for Surf-3 are HMG or ribosomal protein genes (see Huxley et al 1988).

Subsequently to the work presented here, Dr. Agata Giallongo in this laboratory has taken two peptides, one from the carboxy and the other from the amino terminus of the predicted Surf-3 protein. Polyclonal antibodies raised against these peptides detect an abundant 32 kDa protein in a variety of cell extracts. Furthermore, this 32 kDa protein is localized in the cytoplasm and nucleolus as shown by immunoflourescence and can be purified with the 60S ribosomal subunit (Giallongo et al 1989). This indicates that Surf-3 is a ribosomal protein and it has been identified as L7a both by its size and its homology to the previously published 39 amino acid sequence of the N-terminal region of rat L7a as determined by Wittmann-Liebold et al (1979) (see Giallongo et al 1989).

Surf-3 is only the sixth published mammalian intron containing ribosomal protein gene, the others being: mouse S16 (Wagner & Perry 1985), mouse L30 (Wiedemann & Perry 1984), mouse L32 (Dudov & Perry 1984), human S14 (Rhoads et al 1986) and human S17 (Chen & Roufa 1988). A number of cDNAs of other mammalian ribosomal proteins have been isolated but as each gene is a member of a multigene family, it has been difficult to

clone the expressed genes. In every case the mRNA is approximately 1 kb in size and there is a very short 5' untranslated leader and 3' untranslated region.

No detailed analysis of the 5' upstream sequence of a mammalian ribosomal protein gene has been published but various homologies have been observed in the upstream regions of the S16, L30 and L32 gene (Dudov and Perry 1986). These include a poly pyrimidine stretch spanning the CAP site and a sequence ACCNGCTTCYCTY at approximately -70 from the CAP site. Deletion analysis has identified a number of regions which are important for full expression of L32 (Atchison et al 1989). In particular a consensus sequence $C^{G}/_{T}GCCATCT$ has been derived for an element important for high expression which is located in the first exon of L32 (and also c-myc) (Atchison et al 1989). It has also been found that the first 27 bp of the first intron of L32 increases the transcription of this gene by 5 to 10 fold (Chung & Perry 1989). In addition to the mammalian ribosomal protein genes, over twenty *S. cerevisiae* genes have been cloned. Most of the *S. cerevisiae* genes have the consensus DNA sequence, $AACA^{T}/_{C}CC^{G}/_{A}T^{G}/_{A}CAT^{T}/_{C}^{T}/_{A}$, for the binding site of the transcription factor TUF located less than 500 bp 5' to the CAP site.

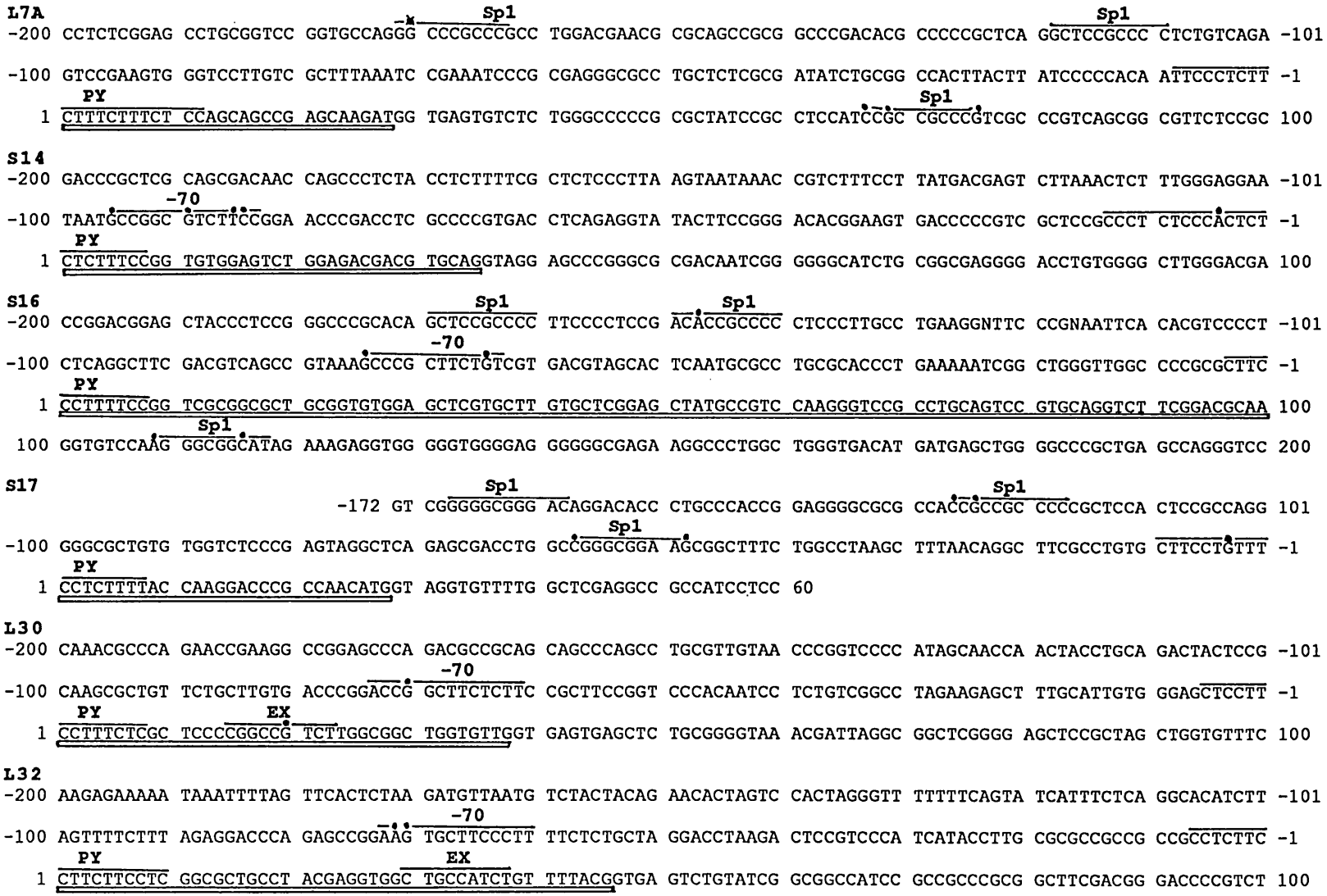
Figure 2.7 shows a comparison of the sequences at the 5' ends of the mouse Surf-3/L7a, human S14, mouse S16, human S17, mouse L30 and mouse L32 genes. As can be seen, the pyrimidine tract spanning the Cap site is prominent in all six promoters. The consensus sequence $C^{G}/_{T}GCCATCT$ is found in the first exons of the L30 and L32 genes but not in the first exons of the other genes. The first 27 bp of the first intron of the L32 gene has no homology to the first introns of the other genes. The consensus sequence ACCNGCTTCYCTY is found at roughly the same position in the S14, S16, L30, L32 genes. The L7a, S16, and S17 genes upstream regions all contain sequences which match the consensus binding site for the transcription factor Sp1

(Kadonaga et al 1986) but the S14, L30 and L32 genes upstream regions do not contain any sequences matching the Sp1 consensus sequence. None of the upstream regions of the mammalian ribosomal protein genes shown in figure 2.7 contain sequences matching the consensus DNA binding site for the *S. cerevisiae* transcription factor TUF, though binding sites could occur further upstream than I have looked. Overall, it appears that apart from the poly pyrimidine tract, none of the recognized consensus sequences are common to all six mammalian ribosomal protein gene 5' regions. The significance of the lack of common features in these 5' regions is discussed in section 2.6.

Figure 2.7 Comparison of sequences at the 5' ends of ribosomal protein genes.

The DNA sequence running from approximately -200 to +100 in relation to the CAP site is shown for each ribosomal gene. The sequences for human S14, mouse S16, human S17, mouse L30 and mouse L32 are taken from Rhoads et al 1986, Wagner & Perry 1985, Chen & Roufa 1988, Wiedemann & Perry 1984 and Dudov & Perry 1984 respectively.

The numbering for each gene starts at the (first) CAP site and the first exon of each gene is underlined twice. The poly pyrimidine tracts spanning the CAP sites are overlined and labelled **PY**. Sequences resembling the consensus sequence $C^{G}/_{T}GCCATCT$ in the first exon of a gene is overlined and labelled **EX**. Sequences resembling the consensus sequence ACCNGCTTCYCTY at approximately -70 are overlined and labelled **-70**. Sequences which contain the 6 invariant nucleotides in the consensus binding site for the transcription factor Sp1, $^{G}/_{T}GGGCGG^{G}/_{A}{}^{G}/_{A}{}^{C}/_{T}$ (Kadonaga et al 1986), are overlined and labelled **Sp1**. For all the sequences matching a consensus, matches are overlined and mismatches marked with a . .


L7A
Sp1
Sp1
-200 CCTCTCGGAG CCTGCGGTCC GGTGCCAGGG CCCGCCCGCC TGGACGAACG CGCAGCCGCG GCCCGACACG CCCCCGCTCA GGCTCCGCCC CTCTGTCAGA -101
-100 GTCCGAAGTG GGTCCTTGTC GCTTTAAATC CGAAATCCCG CGAGGGCGCC TGCTCTCGCG ATATCTGCGG CCACTTACTT ATCCCCCACA ATTCCCTCTT -1
PY
Sp1
1 CTTTCTTTCT CCAGCAGCCG AGCAAGATGG TGAGTGTCTC TGGGCCCCCG CGCTATCCGC CTCCATCCGC CGCCCGTCGC CCGTCAGCGG CGTTCTCCGC 100
S14
-200 GACCCGCTCG CAGCGACAAC CAGCCCTCTA CCTCTTTTCG CTCTCCCTTA AGTAATAAAC CGTCTTTCCT TATGACGAGT CTTAAACTCT TTGGGAGGAA -101
-70
-100 TAATGCCGGC GTCTTCCGGA ACCCGACCTC GCCCCGTGAC CTCAGAGGTA TACTTCCGGG ACACGGAAGT GACCCCCGTC GCTCCGCCCT CTCCCACTCT -1
PY
1 CTCTTTCCGG TGTGGAGTCT GGAGACGACG TGCAGGTAGG AGCCCGGGCG CGACAATCGG GGGGCATCTG CGGCGAGGGG ACCTGTGGGG CTTGGGACGA 100
S16
Sp1
Sp1
-200 CCGGACGGAG CTACCCTCCG GGCCCGCACA GCTCCGCCCC TTCCCCTCCG ACACCGCCCC CTCCCTTGCC TGAAGGNTTC CCGNAATTCA CACGTCCCCT -101
-70
-100 CTCAGGCTTC GACGTCAGCC GTAAAGCCCG CTTCTGTCGT GACGTAGCAC TCAATGCGCC TGCGCACCCT GAAAAATCGG CTGGGTTGGC CCCGCGCTTC -1
PY
1 CCTTTTCCGG TCGCGGCGCT GCGGTGTGGA GCTCGTGCTT GTGCTCGGAG CTATGCCGTC CAAGGGTCCG CCTGCAGTCC GTGCAGGTCT TCGGACGCAA 100
Sp1
100 GGTGTCCAAG GGCGGCATAG AAAGAGGTGG GGGTGGGGAG GGGGGCGAGA AGGCCCTGGC TGGGTGACAT GATGAGCTGG GGCCCGCTGA GCCAGGGTCC 200
S17
Sp1
Sp1
-172 GT CGGGGGCGGG ACAGGACACC CTGCCCACCG GAGGGGCGCG CCACCGCCGC CCCCGCTCCA CTCCGCCAGG 101
Sp1
-100 GGGCGCTGTG TGGTCTCCCG AGTAGGCTCA GAGCGACCTG GCCGGGCGGA AGCGGCTTTC TGGCCTAAGC TTTAACAGGC TTCGCCTGTG CTTCCTGTTT -1
PY
1 CCTCTTTTAC CAAGGACCCG CCAACATGGT AGGTGTTTTG GCTCGAGGCC GCCATCCTCC 60
L30
-200 CAAACGCCCA GAACCGAAGG CCGGAGCCCA GACGCCGCAG CAGCCCAGCC TGCGTTGTAA CCCGGTCCCC ATAGCAACCA ACTACCTGCA GACTACTCCG -101
-70
-100 CAAGCGCTGT TCTGCTTGTG ACCCGGACCG GCTTCTCTTC CGCTTCCGGT CCCACAATCC TCTGTCGGCC TAGAAGAGCT TTGCATTGTG GGAGCTCCTT -1
PY
EX
1 CCTTTCTCGC TCCCCGGCCG TCTTGGCGGC TGGTGTTGGT GAGTGAGCTC TGCGGGGTAA ACGATTAGGC GGCTCGGGG AGCTCCGCTAG CTGGTGTTTC 100
L32
-200 AAGAGAAAAA TAAATTTTAG TTCACTCTAA GATGTTAATG TCTACTACAG AACACTAGTC CACTAGGGTT TTTTTCAGTA TCATTTCTCA GGCACATCTT -101
-70
-100 AGTTTTCTTT AGAGGACCCA GAGCCGGAAG TGCTTCCCTT TTCTCTGCTA GGACCTAAGA CTCCGTCCCA TCATACCTTG CGCGCCGCCG CCGCCTCTTC -1
PY
EX
1 CTTCTTCCTC GGCGCTGCCT ACGAGGTGGC TGCCATCTGT TTTACGGTGA GTCTGTATCG GCGGCCATCC GCCGCCCGCG GCTTCGACGG GACCCCGTCT 100

2.5 Discussion: the TRK-h2 Oncogene.

The *trk* oncogene was first isolated by Martin-Zanca et al (1986) from a human colon carcinoma by transfection of genomic DNA into NIH 3T3 cells. A cDNA with oncogenic potential was isolated and found to encode a fusion protein. The 5' 221 amino acids of the protein are highly homologous to non-muscle tropomyosin which seems to have replaced the extracellular domain of a putative transmembrane receptor with tyrosine kinase activity. A fusion had formed in the DNA of the colon carcinoma such that the 5' exons of tropomyosin were physically linked to the 3' *trk* exons so that a hybrid mRNA was formed by splicing.

More recently the *trk* oncogene has been isolated with the 5' exons of the human L7a (Surf-3) gene instead of the non-muscle tropomyosin, and has been called *trk-2h* (Kozma 1988). The oncogene was again detected by transfection into NIH 3T3 cells but in this case the oncogene was actually generated during one of the early rounds of transfection and is not present in the original human carcinoma cell line. The *trk-2h* oncogene was formed by fusion of the same 3' *trk* exons to the 5' two exons of the human L7a gene. It is not clear whether the exact point of truncation of the *trk* proto-oncogene is important for its activation, or whether any promoter causing high constitutive expression of the proto-oncogene gives rise to transforming activity.

Having isolated the *trk-2h* oncogene Kozma et al (1988) proceeded to isolate a human L7a cDNA from a placental cDNA library. The protein which it encodes is identical to Surf-3 apart from a single amino acid, the Histidine at position 574 of mouse Surf-3 (Fig. 2.2) is an Arginine residue in the Human L7a gene. At the DNA level there is 88% identity between the human and mouse L7a genes in the coding region with almost all the changes being silent with respect to the amino acid coded for. In the 5' and 3' untranslated

regions there is 75% identity between the mouse and human genes. It is interesting to note that whereas the coding genes have been highly conserved during the 80 million years since mice and men separated in evolution, the mouse pseudogenes have diverged to a much greater degree though they only arose some 10-25 million years ago. Surf-3 hybridizes to a multigene family in human DNA as well as mouse DNA and assuming that this family also consists of processed pseudogenes, they must have arisen independently in the human and mouse lineages. This assumes that the pseudogenes are diverging at the rate of unselected DNA.

2.6 Further Investigations on Surf-3

There are 60-80 ribosomal proteins in the eukaryotic ribosome which make up as much as 15% of cellular protein (Mager 1988). Thus it is important for the economy of the cell that they should be expressed in equal quantities and be co-regulated under varying conditions of protein synthesis. In the prokaryote *E. coli* control of ribosomal protein synthesis is at two levels. Transcription of the ribosomal protein operons is coregulated under different growth conditions. For example low levels of amino acids causes repression of transcription by stringent control and this is mediated through a common sequence immediately downstream of the Pribnow box, $C^{C}/_{G}{}^{C}/_{G}{}^{C}/_{G}N^{C}/_{G}CNT$ (Travers 1984). There are also feedback mechanisms whereby free ribosomal proteins, which have not been incorporated into ribosomes, bind to their polycistronic mRNA and block translation. One example of this feedback regulation is the L1 protein repression of the L11 cistron (Thomas & Nomura 1987). This ensures that all the ribosomal proteins are made in the same quantity.

Over twenty *Saccharomyces cerevisiae* ribosomal protein genes have been identified and located scattered throughout the genome. It appears that the major level of control of co-regulation and response to environmental

stimuli, in common with other eukaryotic systems described in the introduction (section 1.2), is mediated by trans-acting protein transcription factors acting through DNA binding sequences located in cis to the genes. All the *S. cerevisiae* ribosomal protein genes, except L3 and S33, contain DNA sequences matching the consensus $AACA^{T}/_{C}CC^{G}/_{A}T^{G}/_{A}CAT^{T}/_{C}T/_{A}$, called the UAS_{rpg} (Planta & Raue 1988), located within 500 bp upsteam of the CAP site. The UAS_{rpg} box has been shown to bind a protein transcription factor called TUF, and a single binding site is sufficient for transcription activation (Woudt et al 1986). It is thought that the TUF transcription factor directs the co-expression of the ribosomal genes under different growth conditions (Mager 1988). However there are *S. cerevisiae* genes which do not have a DNA sequence matching the consensus for the UAS_{rpg} and it is not clear how the transcription of these genes is co-regulated with the genes which do have UAS_{rpg} sequences.

When extra copies of rp genes are introduced into *S. cerevisiae* the level of rp mRNA level is elevated in proportion to the number of gene copies. However the protein level is not elevated in yeast cells with extra ribosomal protein gene copies showing that post-transcriptional control ensures equal production of different ribosomal proteins to match rRNA levels. Repression of splicing has been implicated for the post-transcriptional control of yeast ribosomal protein L32 (Warner et al 1986). However, the most commonly observed type of post transcriptional control in yeast is rapid degradation of excess ribosomal proteins which are not incorporated in ribosomes. Rapid degradation is observed with L25 for example (ElBaradi et al 1986). There appears to be no mechanism for increasing ribosomal protein synthesis if one protein is present in lower amounts than the others. Thus when yeast is depleted for L16 there is a decrease in 60S ribosomal subunits (Rotenberg et al 1988).

Mammalian ribosomal proteins are co-ordinately transcribed at different rates under different conditions. For example, the transcription of S16, L18, and L32 are all decreased 1.5 to 2.0 fold in mouse muscle fibers in comparison to undifferentiated myoblasts (Agrawal and Bowman 1987). However, as described in section 2.4, the promoters of the six mammalian ribosomal protein genes which have so far been published do not appear to have any features in common except the poly pyrimidine tract spanning the CAP site, and no transcription factor which directs co-regulation of transcription has been identified. Clearly it is of interest to identify the factors which are involved in transcription of Surf-3 and to see whether these are the same factors as are involved in the transcription of the other ribosomal protein genes.

Many levels of post transcription control have been observed for vertebrate ribosomal proteins. Post transcriptional control at the level of protein feedback blocking splicing, has been demonstrated for the L1 ribosomal protein gene in Xenopus when this gene is over transcribed (Bozzoni et al 1984). However in similar experiments L14 ribosomal protein was overexpressed and no post transcriptional regulation was observed (Bozzani et al 1984). Translational regulation has been demonstrated in murine lymphosarcoma cells in the presence of the glucocorticoid, dexamethasone (Meyuhas et al 1987). Also, the translational efficiency of S16, L18 and L32 mRNAs is reduced by 1.3 to 1.6 fold in muscle fibers in comparison to undifferentiated myoblasts (Agrawal & Bowman 1987). Post-translational control has been observed when S16 and L32 are over expressed in mouse myoblast cell lines. The mRNA level is proportional to the number of genes but the level of ribosomal protein is unaffected by the level of mRNA due to the rapid degradation of the excess proteins (Bowman 1987). We would like to know if post-transcriptional control occurs with Surf-3 and whether this involves the other genes of the Surfeit locus.

Surf-3 encodes the ribosomal protein L7a, but this does not mean that the Surfeit cluster is a cluster of ribosomal protein genes. Surfs-1, 2 and 4 are not highly expressed and they are unique genes rather than members of multigene families. Furthermore, the ribosomal protein genes which have been mapped are located on different chromosomes, rather than being clustered, as described in chapter 7. However, the other Surfeit genes might be factors involved in translation which are co-regulated with ribosomal proteins. Examples of co-regulation of other proteins with ribosomal proteins are found in *E. coli* and *S. cerevisiae*. In *E. coli*, ribosomal proteins operons contain the genes for RNA polymerase subunits α, β and β' and the elongation factors EF-G and EF-Tu. In yeast the consensus binding sequence for the TUF transcription factor, which causes co-regulation of ribosomal protein genes, is also found upstream of the genes for elongation factor Ef-1α and a common subunit of RNA polymerases I and III.

Finally, the identity of Surf-3 as a ribosomal protein indicates different conditions under which Surf-3 transcription might be differentially regulated. For example, the rate of ribosomal protein synthesis in serum starved 3T3 fibroblasts is half that of serum stimulated cells (Tushinski & Warner 1982). Clearly it is of interest to know whether the other Surfeit genes are co-regulated with Surf-3 under these other conditions, or whether they are expressed independently, and how this relates to the layout of the Surfeit cluster.

Chapter 3: Structural Analysis of Surf-4

3.1 Introduction

Surf-4 lies to the right of the Surfeit Cluster, being transcribed along the genome towards Surf-2 and indeed overlapping with Surf-2 by 133 bp at its 3' end as shown in figure 3.1. Surf-4 has previously been shown to produce two mRNAs of 3.0 kb and 2.4 kb (Williams & Fried 1986b). The sizes of the mRNAs were calculated from their relative positions on a Northern blot with relation to the 18S and 28S rRNAs. More recent Northern blots using RNA size markers of 1.4 kb, 2.4 kb and 4.4 kb (BRL), indicate that the two mRNAs are nearer to 2.0 kb and 2.8 kb in size. Here the larger mRNA will be referred to as the 2.8 kb mRNA and the smaller one as the 2.0 kb mRNA. Only the 2.8 kb mRNA overlaps Surf-2. Two cDNAs called T5B and RMC4, the 3' ends of which overlap Surf-2, and the longest of which was 1.7 kb long, were previously cloned (Williams 1986). Surf-4 is a differentially regulated gene with both the 2.8 kb and 2.0 kb mRNAs being down regulated with respect to Surf-2 in undifferentiated teratocarcinoma cell lines (Williams & Fried 1986b). Whereas in differentiated cell lines, Surf-4 is expressed at roughly 0.05% of the total Poly(A)+ RNA, and Surf-2 is roughly 0.01%, in undifferentiated teratocarcinoma cell lines Surf-4 is expressed at <0.01% (Williams & Fried 1986b).

As discussed in the introduction (section 1.2), it is an extremely rare phenomenon in higher eukaryotic genomes to have overlapping genes. As the mRNAs are transcribed from complementary DNA strands, the mRNAs are complementary at their 3' ends and could form a double stranded form causing antisense interference. In addition, the RNA polymerases transcribing Surf-2 and Surf-4 will be progressing in opposite directions along the genome which could cause interference in particular if the polymerases run over the other genes promoter. In order to investigate the

functional importance of this RNA overlap to the expression of the two genes, it is necessary to know more about Surf-4 with respect to its organization along the genome, whether or not it encodes a protein, whether it is highly transcribed, how stable is the mRNA and above all, whether its expression affects that of Surf-2 and vice-versa.

This chapter shows the overall layout and sequence of Surf-4, the putative protein which it encodes, the relationship between the 2.8 and 2.0 kb mRNAs and the sequence of its 5' upstream region. The level of expression of Surf-4 through the cell cycle is discussed in chapter 6, and its relationship with a CpG rich island is shown in chapter 4.

3.2 Results: Genomic and cDNA Clones of Surf-4

The cDNA RMC4 is clearly not a full length copy of the Surf-4 mRNA being only 1.7 kb long whereas the longer Surf-4 mRNA is 2.8 kb in length (some of this difference in size is due to the presumed Poly(A) tail). In order to obtain full length cDNAs, a cDNA library was generated using Poly(A)+ RNA from Balb/C TS-A-3T3 cells. The cDNA was methylated at the EcoRI sites, then ligated to EcoRI linkers before cloning into lambda NM1149 (a vector very similar to lambda GT10) and screening with RMC4. Out of a library of 30,000, five cDNAs were obtained. One, cDNA 15, was found to be nearly full length being 2.7 kb long. Both cDNA RMC4 and cDNA 15 were sequenced on both strands and found to be exactly homologous over the entire length of RMC4. The cDNA 15 simply contains an extra 828 bp at its 5' end. Unfortunately none of the cDNAs cloned correspond to the shorter 2.0 kb mRNA. Appendix 2b shows the strategy used for sequencing the cDNAs.

The 19 kb EcoRI fragment originally isolated as containing the MES-1 enhancer, Surf-1, Surf-2 and Surf-3 only contains the 3' end of Surf-4. To extend our knowledge of the layout of Surf-4 and the Surfeit cluster as a whole an overlapping genomic fragment was cloned. When the Surf-4 cDNA

RMC4 is used to probe a Southern blot of restriction enzyme digested mouse cellular DNA, it hybridizes to a unique 14 kb BamHI fragment which overlaps with the 19 kb EcoRI fragment by 3.3 kb. To clone this BamHI fragment mouse cellular DNA, cleaved with BamHI, was fractionated by sedimentation in a sucrose gradient. The fraction containing 14 kb fragments was cloned in the lambda vector EMBL 4. Two positives were isolated by screening the resulting library of 110,000 plaques. Both positives contained an insert of 14 kb. The Surf-4 exons contained within this 14 kb BamHI fragment were located by probing southern blots of restriction fragments of the 14 kb BamHI clone with cDNA 15. The exons were shown by DNA sequencing to be exactly homologous to cDNA 15. Furthermore, when unique fragments of cloned DNA were used to probe southern blots of mouse cellular DNA, the restriction enzyme map of the cloned fragment always matched the sizes of the fragments observed in cellular DNA indicating that the cloned 14 kb BamHI fragment is an unrearranged copy of this section of the Surfeit cluster. Appendix 2a shows the extent of the BamHI clone, the subclones derived from it, the exons of Surf-4 which it contains. The sequencing strategy for these exons is shown in appendices 2b and 2c.

The cDNA 15 is homologous to the RMC4 cDNA but is longer at the 5' end. Although RMC4 is completely contained within the 19 kb EcoRI and 14 kb BamHI fragments, cDNA 15 is not. Figure 3.2a shows that the 5' end of cDNA 15 hybridizes to a second BamHI fragment. Figure 3.2b shows that the most distal unique fragment of the 14 kb BamHI clone hybridizes to a 15 kb BclI fragment which overlaps the 14 kb BamHI fragment by 3.5 kb. This BclI fragment was cloned using lambda dash (stratagene) with BamHI ends and size selected BclI restriction fragments of mouse DNA. A library of 150,000 plaques was screened and 2 positives were obtained. Numerous attempts were made to subclone the whole BclI fragment using the flanking SalI sites in the vector, into Bluescript or pxf3 without success. Upon digested of the 15 kb

BclI fragment with EcoRI and SalI three fragments were generated which were successfully cloned into bluescript. One Surf-4 exon was located by southern blotting and shown by sequencing to be exactly homologous to the 5' end of Surf-4 cDNA 15. Furthermore the restriction enzyme map of the fragments corresponds to the bands seen on hybridization to Southern blots of mouse cellular DNA showing that the cloned fragments are unrearranged clones of the distal end of the surfeit cluster. Appendix 2a shows the extent of the BclI fragment, the SalI-EcoRI and EcoRI-EcoRI subclones and the exons of Surf-4. Appendix 2c shows the sequencing strategy for the 5' exon.

The sequencing of the Surf-4 cDNAs and genomic fragments shows that the 2.8 kb Surf-4 mRNA is arranged in six exons spread over 14 kb as shown in figure 3.1. The sequence of Surf-4 is given in figure 3.3 with the positions of the introns marked. The intron sequences are not shown but the splice donor and acceptor sites correspond to the mammalian consensus as derived by Mount (1982). Surf-4 has a very long untranslated 3' region of 1872 bp and this all lies in the 3' most exon. Just before the 133 bp overlap with the 3' end of Surf-2, there is a 63 bp imperfect direct repeat, but there are no significant inverted repeats which would result in hairpin structures. There is a single long open reading frame (ORF) of 269 amino acids starting at the first AUG of the mRNA which is located in a sequence highly homologous to the consensus derived by Kozak (1987) for the vertebrate start of translation. The ORF is spread over all six exons, with the 5' most exon being embedded in a CpG-rich island, see chapter 4. The putative protein encoded by Surf-4 has a pKi of 7.8 and the amino acid composition is almost the same as the average composition of vertebrate proteins (Doolittle 1986). The protein does not appear to have a hydrophobic leader sequence or any other of the well recognized protein homologies and is not homologous to any DNA sequences in the EMBL and GENBANK databases or to any protein sequences in the PIR

database. Appendix 4a shows the protein motifs looked for in the Surf-4 protein.

Figure 3.1. Organization of Surf-4 and the Surfeit cluster

The cloned genomic DNA is shown as a thick bar and each of the arrows above show the extent and direction of each transcription unit in the Surfeit cluster. The CpG-rich islands at the 5' end of each gene (see chapter 4) are shown with solid boxes, and the distance between adjacent islands is indicated. The 19 kb EcoRI, 14 kb BamHI and 15 kb BclI genomic clones are indicated.

Above is an enlargement of Surf-4 showing its intron and exon structure. The six Surf-4 exons are indicated as boxes with the introns being the thinner connecting line, open boxes indicate the open reading frame.

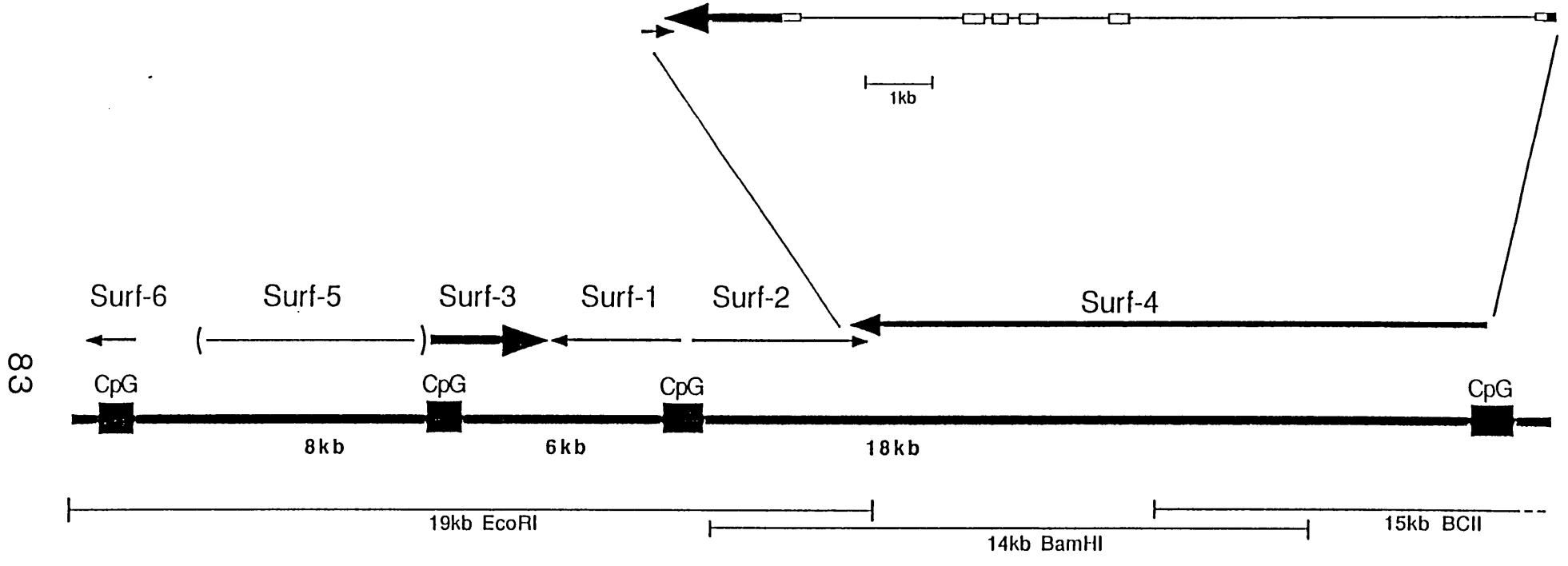

1kb
Surf-6
Surf-5
Surf-3
Surf-1
Surf-2
Surf-4
CpG
CpG
CpG
CpG
8kb
6kb
18kb
19kb EcoRI
14kb BamHI
15kb BCII

Figure 3.2. Southern and Northern blots of Surf-4

a) Southern blot probed with the 400 bp EcoRI to Pst fragment from the 5' end of cDNA 15 (see appendix 2b). Each lane contains 5 µg of Balb/C 3T3 A31 cellular DNA digested with SacI, SalI, HindIII, BamHI or EcoRI. The size markers are HindIII fragments of lambda phage DNA.

b) Southern blot probed with a fragment from the right hand end of the 14 kb BamHI clone called SEB1 PstI (see appendix 2a). Each lane contains 5 µg of Balb/C 3T3 A31 cellular DNA digested with BamHI, BamHI and BclI or BclI alone. The size markers are HindIII fragments of lambda phage DNA.

c) Northern blot showing the 2.0 kb and 2.8 kb Surf-4 mRNAs. A Northern blot of 5 µg of Balb/C TS-A-3T3 poly(A) + RNA was probed with the 928 bp EcoRI fragment of cDNA 15 (see appendix 2a). The position of the RNA size markers (BRL) are indicated on the left.

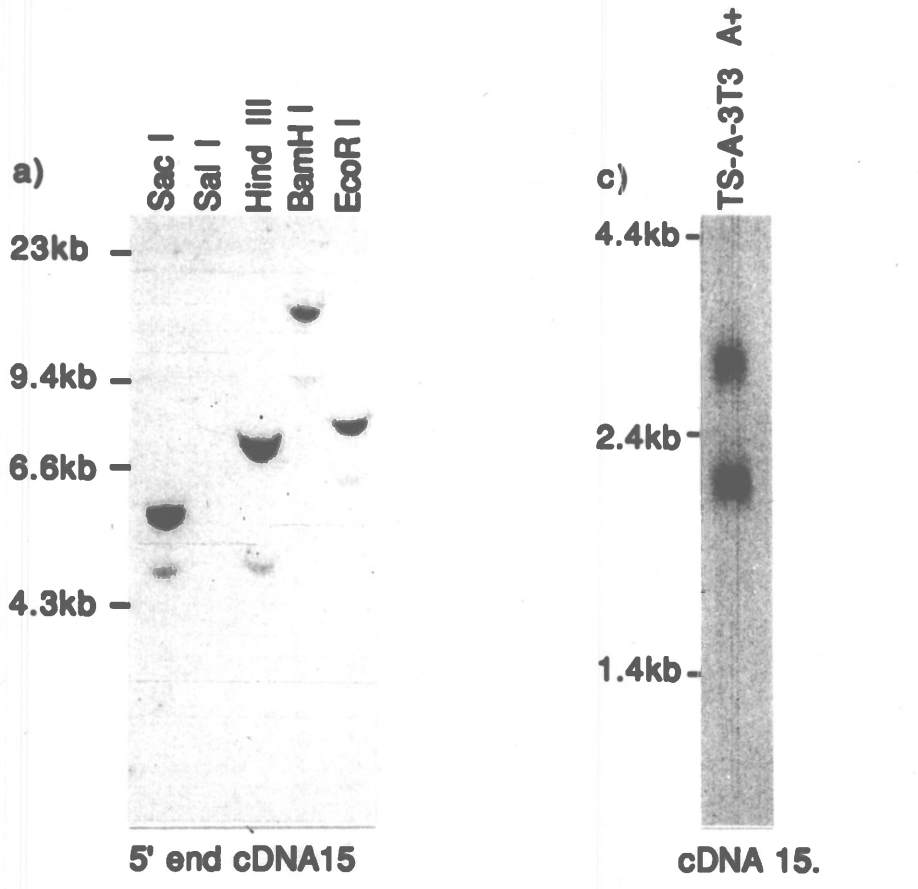

a)
Sac I
Sal I
Hind III
BamH I
EcoR I
23kb
9.4kb
6.6kb
4.3kb
5' end cDNA15
c)
TS-A-3T3 A+
4.4kb
2.4kb
1.4kb
cDNA 15.
b)
BamH I
BamH I + Bcl I
Bcl I
23kb
14kb BamHI fragment
9.4kb
6.6kb
4.3kb
15kb BCII fragment
3.5kb overlap of BamHI and BCII fragments
SEB1 Pst1

Figure 3.3 Sequence of Surf-4.

The numbering starts at the start of transcription as determined by S1 mapping and primer extension (see section 3.3). The 5' end of the longest cDNA, 15, is at position 56 and the 5' end of cDNA RMC4 is at position 885. The protein encoded by the long open reading frame is given starting with the first AUG of the mRNA. The positions of the five introns are indicated although the intron sequence is not shown. The EcoRI sites defining the 928 bp fragment used for S1 mapping to locate the 3' end of the 2.0 kb mRNA are shown at positions 1350 and 2277. The putative poly(A) addition signal at the 3' end of the 2.0 kb mRNA (TTTAAA) and the poly(A) addition signal at the 3' end of the 2.8 kb mRNA (AATAAA) are underlined twice. The 63 bp direct repeat is underlined and the position of the 3' end of Surf-2 is shown.

```
   1 CGCTAGAGCAGCCGCGGCTCAGCGAGGCCGGCCGCCGGGATCTTCCTGTGGAGGCCGCAGCACGCGAGC   69

  70 GCCCAGGACCTGCCAGCTGAGCCTTCCGCCGCCGCCATGGGACAGAACGACCTGATGGGCACGGCCGAG  138
                                        METGlyGlnAsnAspLeuMETGlyThrAlaGlu
                     intron 1
 139 GACTTCGCCGACCAGTTCCTTCGAGTCACCAAGCAGTACCTGCCTCATGTGGCGCGCCTCTGCCTGATC  207
     AspPheAlaAspGlnPheLeuArgValThrLysGlnTyrLeuProHisValAlaArgLeuCysLeuIle

 208 AGCACCTTCCTGGAGGATGGCATCCGCATGTGGTTCCAGTGGAGTGAGCAGCGTGACTATATCGACACC  276
     SerThrPheLeuGluAspGlyIleArgMETTrpPheGlnTrpSerGluGlnArgAspTyrIleAspThr
                                                                  intron 2
 277 ACCTGGAGCTGTGGCTACCTGTTGGCCTCATCCTTCGTGTTCCTCAACCTGCTGGGACAATTGACTGGC  345
     ThrTrpSerCysGlyTyrLeuLeuAlaSerSerPheValPheLeuAsnLeuLeuGlyGlnLeuThrGly

 346 TGTGTTTTGGTGCTGAGCAGGAACTTCGTGCAGTATGCCTGCTTTGGGCTGTTTGGAATCATCGCACTG  414
     CysValLeuValLeuSerArgAsnPheValGlnTyrAlaCysPheGlyLeuPheGlyIleIleAlaLeu
        intron 3                                          intron 4
 415 CAGACAATTGCCTATAGCATTTTGTGGGACTTGAAGTTTCTCATGAGGAACCTGGCTTTAGGAGGAGGT  483
     GlnThrIleAlaTyrSerIleLeuTrpAspLeuLysPheLeuMetArgAsnLeuAlaLeuGlyGlyGly

 484 TTGCTGCTGCTCTTGGCAGAATCCCGCTCAGAAGGGAAGAGCATGTTTGCTGGTGTCCCAACCATGCGT  552
     LeuLeuLeuLeuLeuAlaGluSerArgSerGluGlyLysSerMETPheAlaGlyValProThrMETArg

 553 GAGAGCTCCCCCAAGCAGTACATGCAGCTTGGAGGAAGGGTCCTGCTGGTCCTGATGTTCATGACCCTC  621
     GluSerSerProLysGlnTyrMETGlnLeuGlyGlyArgValLeuLeuValLeuMETPheMETThrLeu
                          intron 5
 622 CTTCACTTTGATGCCAGCTTCTTTTCTATCATCCAGAACATTGTGGGCACAGCTCTGATGATCTTAGTA  690
     LeuHisPheAspAlaSerPhePheSerIleIleGlnAsnIleValGlyThrAlaLeuMETIleLeuVal

 691 GCCATCGGTTTTAAAACCAAGCTGGCAGCTTTGACTCTCGTCGTCTGGCTGTTTGCCATCAACGTGTAT  759
     AlaIleGlyPheLysThrLysLeuAlaAlaLeuThrLeuValValTrpLeuPheAlaIleAsnValTyr

 760 TTCAACGCCTTCTGGACAATCCCGGTATATAAACCCATGCATGACTTCCTGAAATACGACTTCTTCCAG  828
     PheAsnAlaPheTrpThrIleProValTyrLysProMETHisAspPheLeuLysTyrAspPhePheGln

 829 ACCATGTCAGTGATTGGAGGCCTGCTCCTGGTGGTGGCTCTTGGCCCAGGGGGTGTGTCCATGGATGAG  897
     ThrMETSerValIleGlyGlyLeuLeuLeuValValAlaLeuGlyProGlyGlyValSerMETAspGlu

 898 AAGAAGAAAGAGTGGTAACACACAGATCCCTCCCCTCCCTGGCTGAGGCACAGGGCCCTGGCCTGGTTC  966
     LysLysLysGluTrp

 967 AGGGCAGAGTCAACAAACTGCCGGCGTTTGTGTGTCCTTCTCCCTTCCCCTCCCTTGGTAAAGGCACAG 1035

1036 ATGTTTTGAGAACTTTATTTGCAGACACCTGAAAATTGGAGATAAAATCTTTGGAAATAGTCTGGAGTC 1104

1105 TTGACTGTCCAGGGCTGGCGAGCTGGATGGTCACTCCTTAGCCAAGGCTTGGAGGAGACAGCGTGCTGG 1173

1174 CTGTGGCCTCATTCCTCCTGTCTCCGAGTTCCTTTTGGGGAAGTCGTACTGAGCTGTATTTTTATCATT 1242

1243 CAAAGCAGTTTCCCTTCTTAGTTTTGAGTTTACATCCTTAGCTCAAACTACTAAATGACTTAGGATAGT 1311
                                        EcoRI
1312 CCAGCCAAACACCAAACGGTTAAACCTCCAGTTCAGAATTCTACCAGTGGCCCCAGCATTGCTTCTGCC 1380

1381 CCTGTATTTTATAGGATCAGTGAAGATGTTAGAGAAACCAAGTCAAGAGAACAGTACAGTATTGGAAAG 1449

1450 GTGGGTGTTACTGCTGTCCCAAGTCCTTGATGTTAGCTAGCTGGATGCTGTTTACTGGGTGAAAGGTAC 1518

1519 CTTGGGAGTTTTCATGGCTCCAAGCAGGGATGGGATGGGGTAAGCAGCAGGAAGTTCTTAGAAAGCTGC 1587

1588 TATCCCTCAAGCTTGGTGGTGGAAAGGGCAGGAGAGCAGGCTGCTCACAGGATGGCCCACACAACCACC 1656

1657 ATGGAACCAGCCTGGGAACAGCGATGCCATGGGGCCCAGAAGTGAGGAGTCCCAGTACTCAAGACACCC 1725
```

1726 AGGCTCAACCCAGTTACCTCACACTTGACCTTCTCAGAGCAAGGCCTGAACTCCACAGGTTGTGGTCCC 1794

1795 CAACCCCCAGCTTAACTCCCTGTTCCAGGCTGAGCCGACCCCGTCGACAGCCCTGATACCACCATTCCC 1863

1864 CTGAGGCTGTTGCTTGCTCAGATTTTTGTCAACTTGCTCTGCTGTATGCAGCCAAGATACTTTTTACAA 1932

1933 TTTGTGATGCCTTACCAATTTGATCTCAGTCCTGTATTTAAAGTTTTCTAACATTGCCTTACACCGTGT 2001

2002 TTCTCTTTTTGGGGGAGGGGTGGGGTGTCTAATGCTGTGGTCTTGCTCATCTGCATGGTAGTGCCATGA 2070

2071 GGGTAGTTGGGTTCCTCCTGGATCTCTCCCAGGTCTGTTGGAGAGAACAGACAGACTTAGAACCCCTGG 2139

2140 AGGGAAGCTGGCAGCTTACCTCACCCTTGGGTACAGAGTCTTGGAGGTCCTCCTTTCTGGCATCAGCTG 2208

EcoRI

2209 GGTAGGCAGGACTTTGTTCTCTGAGTCACTTTGGGTCAGTCCCTTGCCTCGCTGGCTCTGCAGGCTGAA 2277

2278 TTCACCTTCTCAACACCTAGCCCATGTTTTCCTGCAGCCTTATGGCCTGCCTATTCCTCCCAAGTGAGT 2346

2347 TCCAGCACTGGCTCAGGCCATGCTAGTTCTTTACCACAGAGCTGCCTCAGTGCACAGCCACCTGCAGTG 2415

2416 GCAGCCCCATGACATTCCAGTCTGTGGAGTGGGGCAGACCCTGGGAGGCAAGTGCATGCCGTGCAGCAG 2484

2485 AAGCCCTCGTCCCAGGGTTACTTAGCCCTTTGACCAGCCTCCTGTGGGTTCTGCCTGGGATGGGAACCC 2553

2554 TCATCCCAAGGGTACTTGGCCCATTTACCAGCCTCCTGTGGGGTGCGCCTGGGACGGGAGAAAACAATG 2622

3' end of Surf-2

2623 ATACACATATTGGTTTGGCAAAAGTTTTCTGAAGATGCTTTAGCTTTATGTGGGAATAGGTATAAACCT 2691

2692 CATTTTTATGCTGTATTTTATATCTTAGTTGTGTTTGAAAATGTTTTGATCTTTGGAAACATCAAAATA 2760

2761 AATAATGGCATTTGTGGTATATA 2783

\poly(A) addition site

3.3 Results: Locating the 5' End of Surf-4

The 5' end of cDNA 15 is very CpG-rich and contains an initiation ATG triplet in a good environment for initiation of translation as defined by Kozak (1987) suggesting that this cDNA is very nearly full length. To determine the exact position of the 5' end of the mRNA, primer extension and S1 nuclease experiments were carried out. For primer extension a 30 bp deoxyoligonucleotide complementary to a sequence near the 5' end of cDNA 15 (Fig. 3.5) was end labelled and hybridized to Balb/C TS-A-3T3 poly(A)+ RNA before reverse transcription. The result is shown in figure 3.4 and shows that the major start of transcription is 103 bp from the 5' end of the primer.

Because Primer extension can give spurious results due to premature termination of the reverse transcriptase at regions of secondary structure such as are found in CpG-rich sequences, S1 mapping of the 5' end of Surf-4 was also carried out. A 328 bp fragment running from an Nco1 site in the 5' most exon, to an upstream Xba1 site (Fig. 3.5), was end labelled at the Nco1 site and used for S1 mapping as shown in figure 3.4. Multiple start sites are observed spread over some distance and the major site is situated 108 bp from the 5' end of the probe. When the positions of the major start site observed by primer extension and S1 mapping are superimposed on the genomic sequence upstream of the 5' end of cDNA 15, they are found to correspond as shown in figure 3.5.

The genomic sequence upstream of this major start site is very CpG-rich and forms the CpG-rich island located by the presence of island specific restriction enzyme cutting sites in cellular DNA as shown in chapter 4. The 500 bp upstream of the major start site, which probably forms the promoter region, is shown in figure 3.5, it contains five SP1 binding sites, one AP1 binding site, one AP2 binding site and one AP3 binding site as determined by

homology to the published consensus binding sites for these transcription factors. Appendix 4b gives a list of all the factor binding sites looked for. There is no obvious homology between the 5' region of Surf-4 and the 5' region between Surf-1 and Surf-2 and the 5' region of Surf-3. In particular, the Surf-2/Surf-1 region contains no sequences which correspond to the SP1, AP1, AP2 or AP3 consensus binding sites but does contain a MLTF and a CREB site (see appendix 4b for the consensus binding sites of these transcription factors). The 5' region of Surf-3 contains one SP1 site but no AP1, AP2, AP3 or MLTF or CREB sites, it also contains a pyrimidine tract covering the CAP site which is characteristic of ribosomal protein genes (Dudov & Perry 1986).

Figure 3.4 Primer extension and S1 mapping to determine the 5' end of Surf-4.

a) S1 mapping was carried out with a 328 bp NcoI to XbaI end labelled fragment (see Fig. 3.5). The first lane shows the protected fragments after digestion with S1 nuclease of the fragment hybridized to 40 µg of Yeast tRNA, the second lane is the same fragment hybridized to 5 µg of Balb/C TS-A-3T3 poly(A)+ RNA. The righthand lanes contain a sequencing ladder generated using ssM13 DNA with the sizes shown at the right.

b) Primer extension was carried out with the 30 bp oligonucleotide indicated in figure 3.5 as the primer. The righthand lane shows the extended products after the primer was hybridized to 40 µg of yeast tRNA and the nextdoor lane the results using 5 µg of poly(A)+ RNA from Balb/C TS-A-3T3 cells. The lefthand lanes contain a sequencing ladder generated using ssM13 DNA and the sizes are shown on the left.

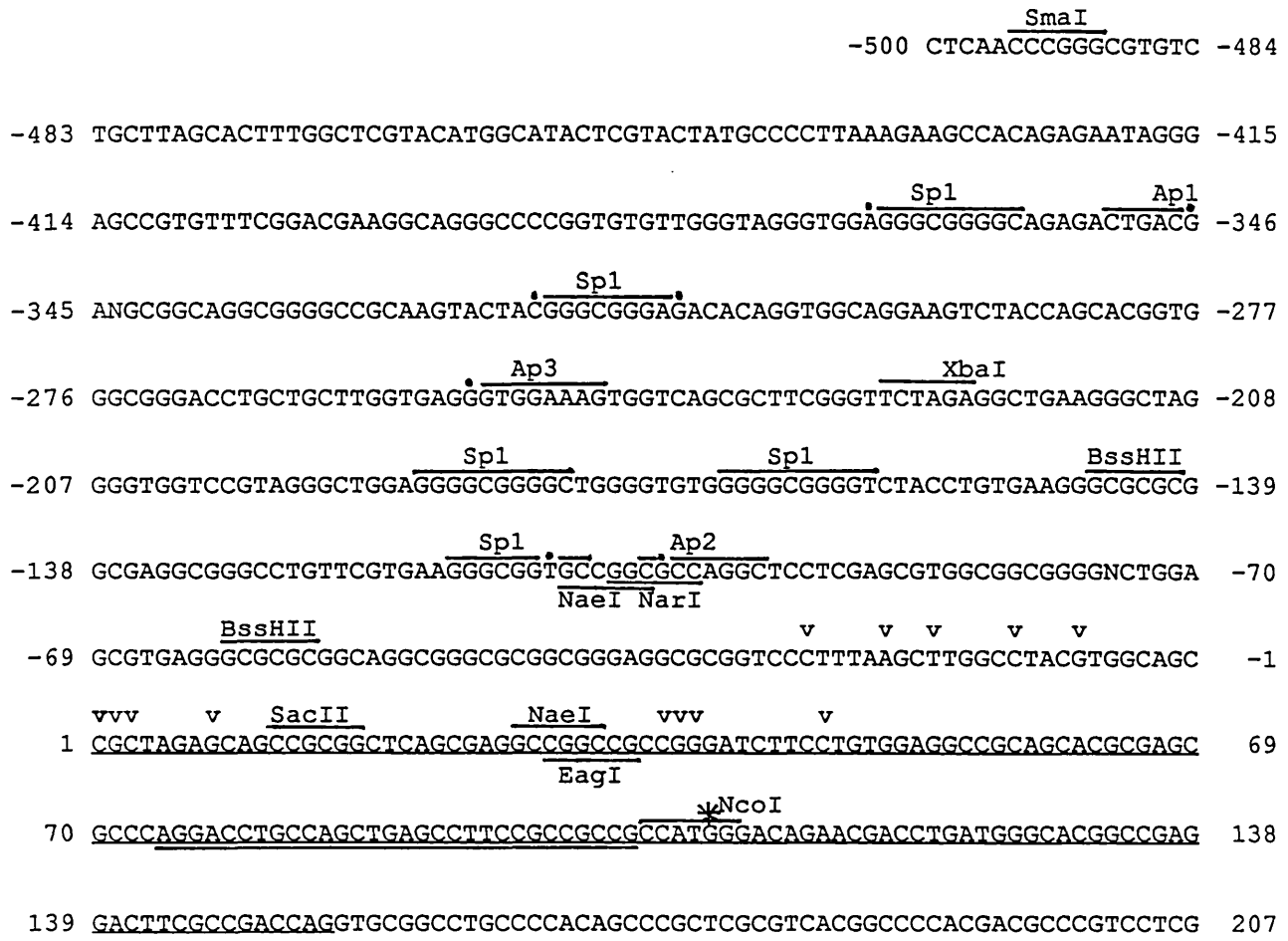
Yeast tRNA
TS-A-3T3
M13 sequence
108-
130
120
110
100
90
80
70
S1 mapping
M13 sequence
TS-A-3T3
Yeast tRNA
120
110
100
90
80
-103
Primer Extension

Figure 3.5 DNA sequence of the upstream region of Surf-4.

The first exon of Surf-4 is underlined and the major start of transcription is numbered 1. The 30 bp oligonucleotide (nt 74-103) used for primer extension is doubly underlined and the * (nt 108) marks the 5' end of the NcoI to XbaI fragment used for S1 mapping. The NcoI and XbaI sites are also indicated. The v's indicate the major and minor start sites.

Sequences which correspond to the consensus binding DNA sequences for various transcription factors are indicated (Sp1, AP1, AP2 and AP3). The consensus DNA sequences are; Sp1 5' $^{G}/_{T}$GGGCGG$^{G}/_{A}{}^{G}/_{A}{}^{C}/_{T}$ 3' (Kadonaga et al 1986), AP1 5' $^{C}/_{G}$TGACT$^{C}/_{A}$A 3' (Lee et al 1987), AP2 5' CC$^{C}/_{G}$C$^{A}/_{G}$GGC 3' (Mitchell et al 1987), and AP-3 5' TGTGG$^{A}/_{T}{}^{A}/_{T}{}^{A}/_{T}$G 3' (Sassone-Corsi & Borelli 1986). Identities with the consensus are overlined and non-matches are marked with a . .

The recognition sites for the island specific rare cutting enzymes which occur in this region are indicated (see chapter 4).

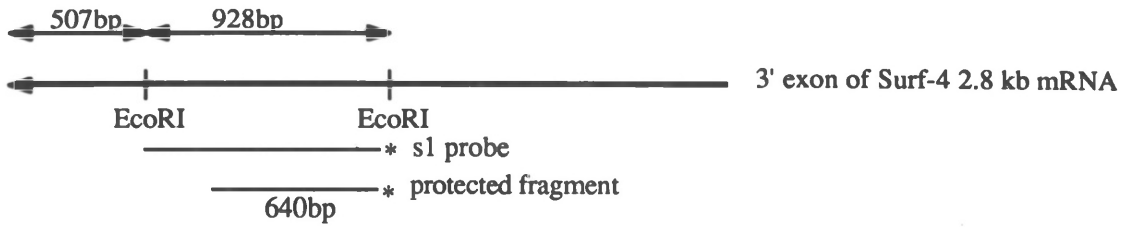

SmaI
-500 CTCAACCCGGGCGTGTC -484
-483 TGCTTAGCACTTTGGCTCGTACATGGCATACTCGTACTATGCCCCTTAAAGAAGCCACAGAGAATAGGG -415
Sp1
Ap1
-414 AGCCGTGTTTCGGACGAAGGCAGGGCCCCGGTGTGTTGGGTAGGGTGGAGGGCGGGGCAGAGACTGACG -346
Sp1
-345 ANGCGGCAGGCGGGGCCGCAAGTACTACGGGCGGGAGACACAGGTGGCAGGAAGTCTACCAGCACGGTG -277
Ap3
XbaI
-276 GGCGGGACCTGCTGCTTGGTGAGGGTGGAAAGTGGTCAGCGCTTCGGGTTCTAGAGGCTGAAGGGCTAG -208
Sp1
Sp1
BssHII
-207 GGGTGGTCCGTAGGGCTGGAGGGGCGGGGCTGGGGTGTGGGGGCGGGGTCTACCTGTGAAGGGCGCGCG -139
Sp1
Ap2
-138 GCGAGGCGGGCCTGTTCGTGAAGGGCGGTGCCGGCGCCAGGCTCCTCGAGCGTGGCGGCGGGGNCTGGA -70
NaeI NarI
BssHII
v v v v v
-69 GCGTGAGGGCGCGCGGCAGGCGGGCGCGGCGGGAGGCGCGGTCCCTTTAAGCTTGGCCTACGTGGCAGC -1
vvv v SacII NaeI vvv v
1 CGCTAGAGCAGCCGCGGCTCAGCGAGGCCGGCCGCCGGGATCTTCCTGTGGAGGCCGCAGCACGCGAGC 69
EagI
NcoI
70 GCCCAGGACCTGCCAGCTGAGCCTTCCGCCGCCGCCATGGGACAGAACGACCTGATGGGCACGGCCGAG 138
139 GACTTCGCCGACCAGGTGCGGCCTGCCCCACAGCCCGCTCGCGTCACGGCCCCACGACGCCCGTCCTCG 207

3.4 Results: The Two Surf-4 Messages Differ in Their Poly(A) Addition Site

None of the cDNAs isolated either in this work or by Trevor Williams (Williams 1986) showed a difference in structure which could easily account for the 2.0 as opposed to the 2.8 kb mRNA. Northern blots showed that whereas the 3' most 507 bp of cDNA RMC4 detects only the 2.8 kb message, the next 928 bp, which can be isolated as a 928 bp EcoRI fragment (see fig. 3.3), detects both messages. A single stranded end labelled probe was prepared from the 928 bp EcoRI fragment which would hybridize to the mRNA and used in an S1 nuclease experiment. As is shown in figure 3.6, this 928 bp single stranded probe gives two protected fragments, one of 928 bp and one of 640 bp. The full length 928 bp fragment is protected by the 2.8 kb mRNA and the smaller fragment is consistent with the 2.0 kb mRNAs having a poly(A) site in the middle of the 928 bp fragment. This poly (A) addition site would correspond to a position in the Surf-4 sequence which is 16 bp 3' to the hexanucleotide TTTAAA (indicated in figure 3.3 and figure 3.6) which could be acting as a poly(A) addition signal. This result strongly suggests that the difference in size between the 2.8 kb and 2.0 kb Surf-4 mRNAs is the result of different sites of polyadenylation at their 3' ends.

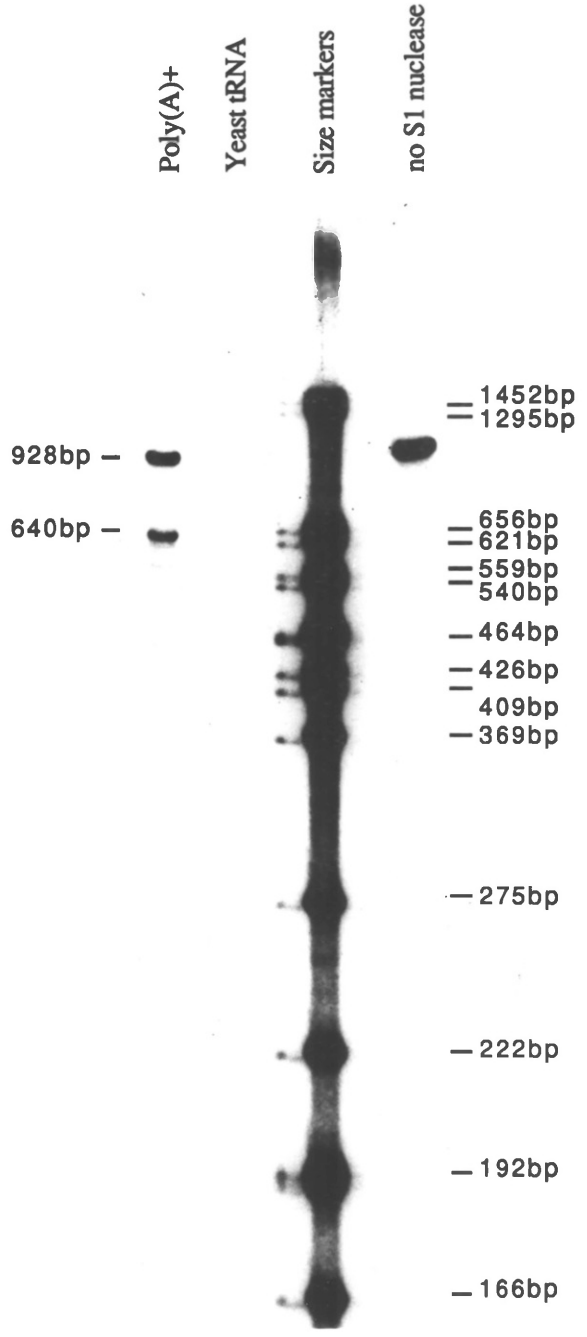
Figure 3.6 S1 mapping to determine the nature of the difference between the 2.0 and 2.8 kb Surf-4 mRNAs.

Above is shown the 3' exon of Surf-4 drawn in the same orientation as maps of the whole Surfeit locus. The 507 bp fragment which only detects the 2.8 kb mRNA and the 928 bp EcoRI fragment used to make the single stranded probe for S1 mapping are shown. Underneath is shown the sequence containing the TTTAAA hexanucleotide (double underline) which might be acting as a poly(A) addition signal for the 2.0 kb mRNA just upstream from the end of the detected 640 bp protected fragment. At the bottom is shown the S1 mapping experiment. The first lane shows the protected fragments after S1 nuclease digestion of the 928 bp fragment hybridized to 5 μg of Balb/C TS-A-3T3 poly(A)+ RNA. The second lane is the result when 40 μg of yeast tRNA is used instead of mouse RNA. The third and fourth lane show size markers which are end labelled fragments of a plasmid the sizes of which are shown on the right. The righthand lane shows the full length probe with no added S1 nuclease.



CCAATTTGATCTCAGTCCTGTATTTAAAGTTTTCTAACATTGCCTTACACCGTGTTTCTCTTTTT

640bp from EcoRI site



3.5 Conclusions and Discussion

The Surf-4 gene is spread over six exons spanning 14 kb on the right hand side of the Surfeit cluster. This size is in marked contrast to the other characterized Surfeit genes which span less than 3 kb of genomic DNA. The gene is transcribed as two mRNAs, of 2.0 and 2.8 kb which differ in their poly(A) addition site, but both messages appear to contain the same open reading frame of 269 amino acids. This putative Surf-4 protein has no homology to a known protein (including the other Surf proteins as determined by dot matrix analysis) and no obvious features. The untranslated 3' 1872 bp of the 2.8 kb mRNA fall into a single exon with the 3' most 133 bp overlapping the 3' end of Surf-2 allowing the possibility of antisense regulation between these two genes. As the shorter 2.0 kb mRNA does not overlap Surf-2 this mRNA would probably not be affected by antisense regulation.

The overlapping of the 3 ends of Surf-2 and Surf-4 means that the 3' end of each gene contains the AATAAA poly(A) addition signal of the other gene and in addition this region is in general A+T rich. Conserved AT-rich sequences have been identified in the 3' untranslated regions of a variety of short lived mRNAs including GM-CSF, tumor necrosis factor, c-myc and c-fos (Caput et al 1986, Shaw & Kamen 1986, Cosman 1987). The consensus sequence TTATTTAT has been identified and a DNA fragment containing this sequence causes low stability when inserted into the 3' untranslated region of the normally stable rabbit β-globin mRNAs (Shaw & Kamen 1986). Similar sequence have also been implicated in rapid turn over of bovine inerleukin 2 mRNA (Reeves et al 1987) and in translation inhibition of human interferon-β mRNA (Kruys et al 1988). Results with c-fos suggest that the 3' AU-rich sequences act to destabilize the mRNA by directing translation dependent rapid removal of the mRNA poly(A) tract (Wilson & Treisman 1988). The

consensus sequence does not occur in the Surf-4 mRNA but it does occur in the Surf-2 mRNA at position 2766 in figure 3.3 so that Surf-2 is possibly a short lived mRNA with control of gene expression occurring post transcriptionally.

The 5' end of the gene lies in a CpG-rich island which is further discussed in chapter 4, and the upstream DNA, presumed to be the promoter region contains a variety of transcription factor consensus binding sites. Although this upstream region is a CpG-rich region like the upstream regions of the other Surfeit genes, there are no other obvious similarities between these regions and they do not contain the same transcription factor binding sites.

Surf-4 is downregulated in undifferentiated teratocarcinoma cells (Williams & Fried 1986b) and it would be interesting to know what sequences in the gene specify this differential expression. Antibodies can be generated against the Surf-4 protein predicted from the open reading frame, and this may allow the identification of Surf-4.

Chapter 4: CpG-Rich Islands

4.1 Introduction

It had previously been noted that the promoter and enhancer region at the 5' end of Surf-1 and Surf-2 is a CpG-rich island, having a higher CpG content than the surrounding DNA, having the expected CpG content for the base composition of the region and being undermethylated (Williams 1986). On sequencing through the 5' region of Surf-3 it became clear that this region too is a CpG-rich island (see section 2.4). It is extremely rare in mammalian genomes to have CpG-rich islands separated by just 6 kb (see section 1.2), so it was of interest to determine whether there were other CpG-rich islands nearby. In particular if Surf-4 also had an associated CpG-rich island this would give an indication of the position of the 5' end of this gene.

As described in the introduction (section 1.2), there are island specific rare cutting restriction enzymes which can be used to locate CpG-rich islands. Table 1.1 shows the restriction enzymes and to what extent the cutting sites are restricted to CpG-rich islands. For the following analysis I have used BssHII, EagI, NaeI, NarI, SacII and SmaI. To locate islands by the presence of cutting sites for these enzymes, genomic DNA rather than DNA cloned in *E. coli* is used as this has the methylation pattern characteristic of mammalian DNA. Furthermore, when cell lines are derived from rodent fibroblasts or cancerous tissue the methylation pattern is often quite different to that found in normal tissues (Wilson & Jones 1983). For example CpG-rich islands which are never methylated in primary tissue can be found methylated and sites outside of islands can be found unmethylated in such cell lines. Indeed, this labile pattern of methylation is apparently responsible for the very high rate of mutation in cell lines derived from the Chinese Hamster Ovary, CHO, cell line. The spleen of mice is a good source of "correctly methylated" DNA, whereas a rodent fibroblast cell line

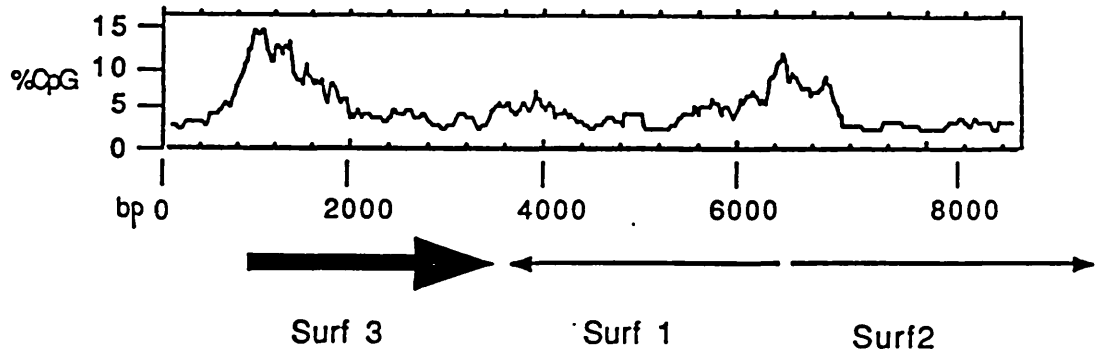
transformed with polyoma virus, such as Balb/C TS-A-3T3 is not. In a number of the following results I have used DNA from the Balb/C 3T3 A31 cell line which is probably not as good as spleen DNA in this respect.

4.2 Results: Four CpG-rich Islands in the Surfeit Cluster

Having sequenced through Surf-3 we now have continuous mouse sequence running from 1.0 kb 5' of Surf-3 to the middle of Surf-2. When the CpG dinucleotide content of the DNA is plotted against distance along the DNA, the islands at the 5' ends of Surf-3 and Surf-1/-2 are quite clear (Fig. 4.1)

Figure 4.1 CpG dinucleotide content of the Surfeit Locus

The percentage of CpG dinucleotides is plotted against bp of DNA sequence. The extent and location of the Surf-1, 2 and 3 transcription units are indicated as arrows.



In order to detect neighbouring islands, southern blots were made of Balb/C 3T3 A31 mouse DNA restricted with island specific rare cutting enzymes. When such a blot is probed with Surf-1 cDNA (Fig. 4.2), a SacII fragment of 6.5 kb is seen indicating that the SacII sites identified in the region between the 5' ends of Surf-1 and Surf-2 and in the region 5' to Surf-3, by sequence analysis, are indeed present and unmethylated. There is also

a large BssHII fragment whose size cannot be measured on this blot. When the blot is probed with a fragment from Surf-4 (probe B in Fig. 4.2), a SacII fragment and a SmaI fragment, both of about 18 kb as well as the same large BssHII fragment seen with the Surf-1 probe, are seen. This suggests the presence of an island at the 5' end of Surf-4 lying about 18 kb from the central island. When it is probed with a fragment from upstream of Surf-3 (probe A in Fig. 4.2) a SacII fragment of 8 kb is seen as well as the same large BssHII fragment, indicating the presence of yet another island 8 kb to the left of Surf-3.

To confirm the island at the 5' end of Surf-4, a southern blot was made with genomic DNA digested with KpnI, and with double digestions of KpnI and island specific rare cutting enzymes. This blot was probed with a DNA fragment located between the KpnI site and the CpG-rich island (probe B in Fig. 4.3). All the CpG-rich island specific enzymes cut the large KpnI fragment within about 500 bp of each other indicating the presence of an island about 6 kb beyond the KpnI site (Fig. 4.3). This island is located at the 5' end of Surf-4 as the genomic DNA covering this region has now been cloned and sequenced and the cutting sites of these rare restriction enzymes in the cloned DNA are indeed clustered around the 5' end of Surf-4 (see Fig. 3.5).

To confirm the island 8 kb to the left of Surf-3 a southern blot was made with genomic DNA digested with BamHI and with double digests of BamHI and the island specific rare cutting enzymes. This blot was probed with a fragment from the far left of the 19 kb EcoRI fragment (probe A in Fig. 4.3). The large BamHI fragment is cut by SacII and BssHII confirming the presence of a CpG-rich island at this position.

The DNA around all four islands has now been sequenced and their nucleotide contents are indeed characteristic of CpG-rich islands. As can be seen in table 4.2, when a 200 bp window in the middle of each island is analyzed, it has a high C and G content and the CpG dinucleotide content is

not repressed. For comparison a 200 bp window from an intron of Surf-3 was also analyzed and found to have a much lower C and G content. The CpG dinucleotide content is considerably reduced in comparison with that expected from the nucleotide content, whereas the GpC content is not significantly lower than the expected value (see Normal in Table 4.1).

Table 4.1 C and G content of the four Surfeit CpG-rich islands.

A window of 200 bp was taken from the islands upstream of Surf-3, and at the 5' ends of Surf-3, Surfs-1 and 2, and Surf-4 respectively. "Normal" values are calculated from a 200 bp window in an intron of Surf-3. "Enzyme sites" indicates the island specific rare enzyme cutting sites located in each island. "%C+G" is the percentage of C and G nucleotides in each 200 bp window. "%CpG obs." is the percentage of CpG dinucleotides in each 200 bp window whereas "%CpG exp." is the percentage of CpG dinucleotides expected from the %C+G. "%GpC obs." and "%GpC exp." are the corresponding values for the dinucleotide GpC.

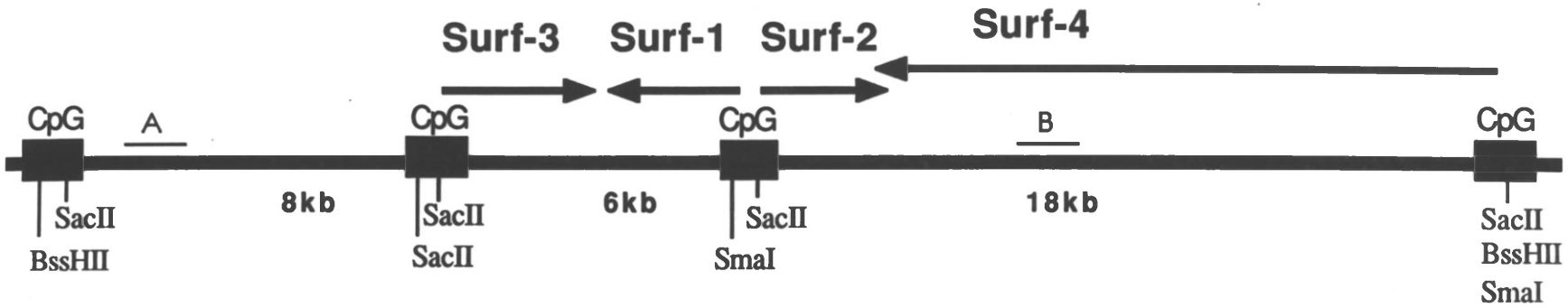
Island	enzyme sites	%C+G	%CpG obs	%CpG exp	%GpC obs	%GpC exp
Upstream	SacII BssHII	63.0	7.5	9.8	11.6	9.8
Surf-3	SacII NarI	66.5	9.9	10.8	9.9	10.8
Surf-1, -2	SacII SmaI NaeI	63.0	9.7	9.5	9.3	9.5
Surf-4	SacII BssHII EagI NaeI NarI SmaI	77.2	13.0	12.8	16.0	12.8
Normal		48.8	1.6	5.5	5.1	5.5

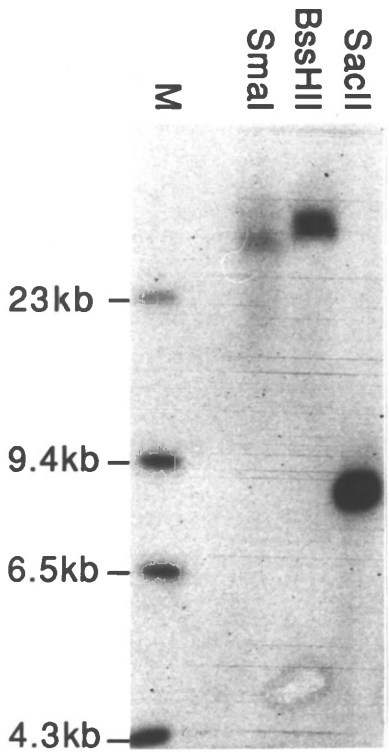
When all the sequence available for the Surfeit cluster was searched for rare cutting enzyme sites a further BssHII cleavage site was found. The BssHII site lies in between the Surf-2 and Surf-4 islands. This site can be cleaved in the cloned DNA but the site in cellular DNA is resistant to cleavage by BssHII indicating that this site is methylated (Fig. 4.4). This confirms that

island specific rare enzyme recognition sites in cellular DNA are largely confined to islands, and those recognition sites which occur in non-island DNA are frequently methylated and not cleaved by the enzyme.

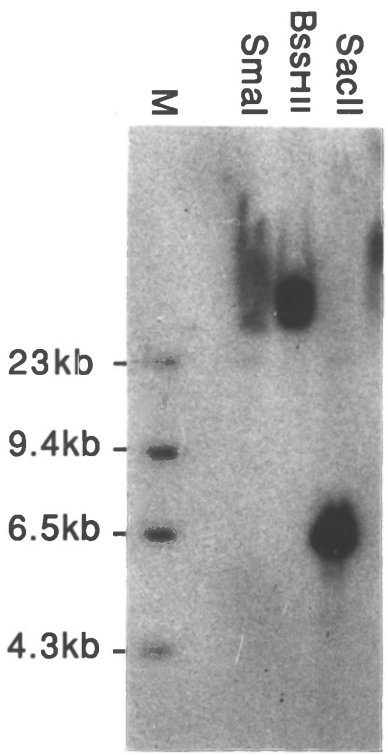
Figure 4.2 Location of four CpG-rich islands in the Surfeit cluster

The Surfeit cluster is diagrammed at the top with the Surf-1 to 4 transcription units shown as arrows. The 4 CpG-rich islands are shown as solid boxes with the locations of the island specific rare enzyme cutting sites indicated. Underneath are three southern blots of mouse cellular DNA cut with SmaI, BssHII or SacII. Each lane contains 5 µg of Balb/C 3T3 A31 DNA. The left hand lane shows lambda HindIII fragments used as size markers. Probe A is a 1.1 kb BglII genomic fragment labelled in appendix 3. The Surf-1 probe was cDNA T1B, and probe B is the 928 bp EcoRI fragment from Surf-4 labelled in appendix 2a.

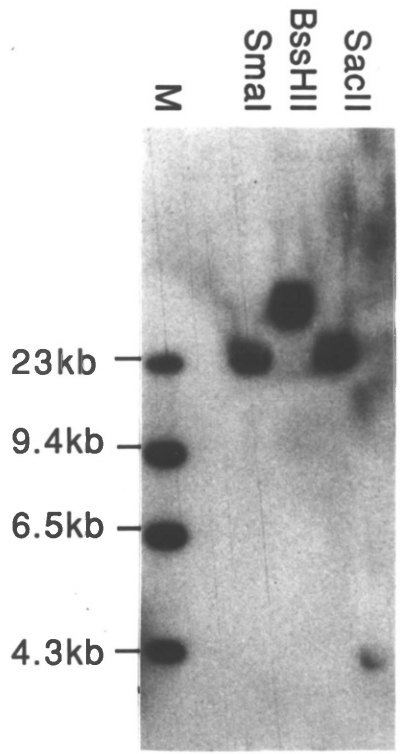




Probe A



Surf-1 Probe



Probe B

Figure 4.3 Further analysis of the CpG-rich islands at the far left of the Surfeit cluster and at the 5' end of Surf-4

The Surfeit cluster is shown at the top with the Surf genes shown as arrows. The CpG-rich islands are shown as solid boxes with the rare enzyme cutting sites shown under each CpG-rich island. The left hand southern blot has mouse cellular DNA digested with the enzymes shown above each lane and is hybridized with Probe A. The position of probe A is shown on the map above and is the XhoI to HindIII fragment labelled as such in appendix 3. The right hand southern blot shows mouse cellular DNA digested with the enzymes shown above each lane and is hybridized with probe B. The position of probe B is shown on the map above and is the fragment labelled SEB 1 Pst in appendix 2a. Each lane of the Southern blots contains 2 µg of Balb/C 3T3 A31 DNA and the positions and sizes of the lambda HindIII fragments used as size markers are indicated.

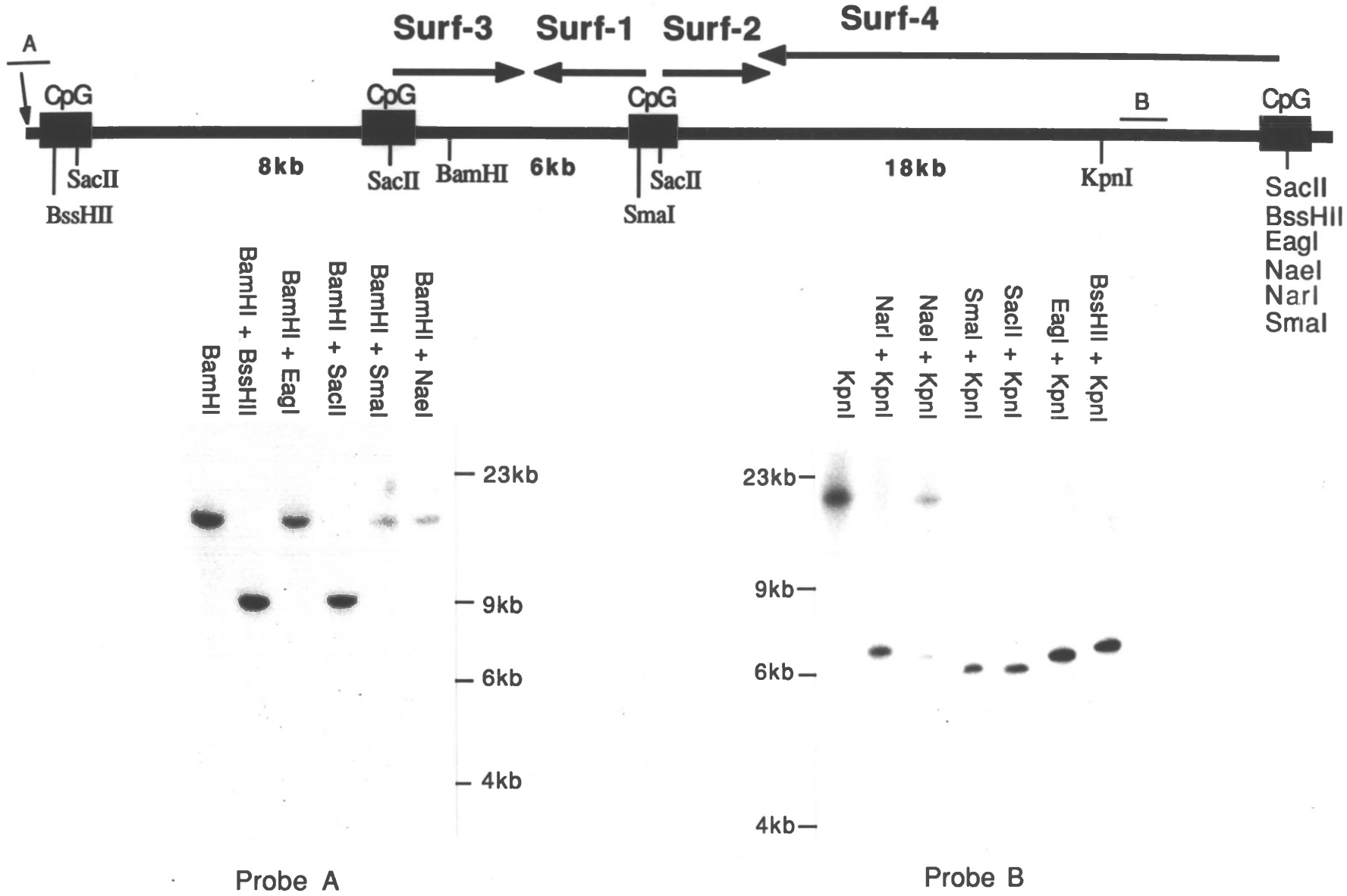
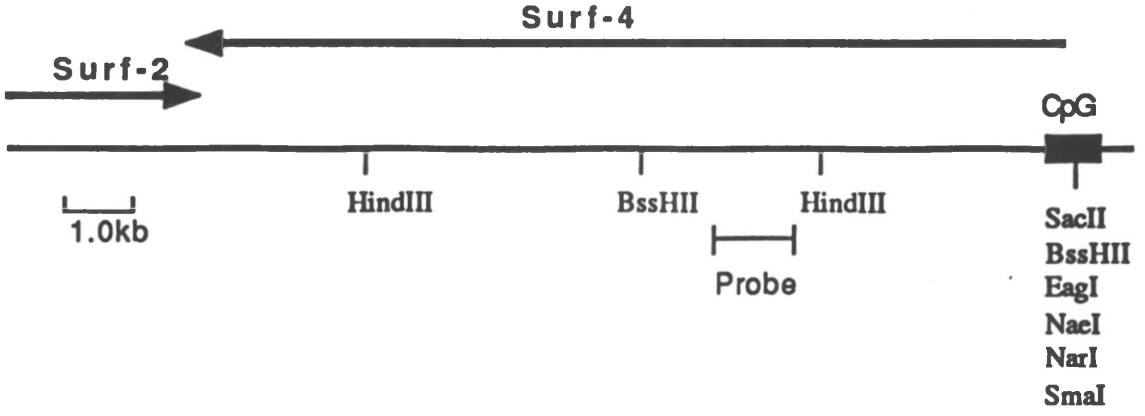
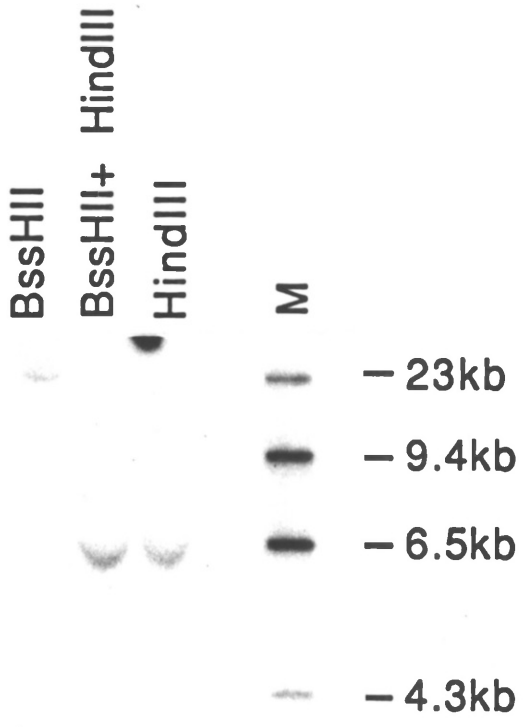

Surf-3
Surf-1
Surf-2
Surf-4
A
CpG
CpG
CpG
B
CpG
SacII
BssHII
8kb
SacII
BamHI
6kb
SacII
SmaI
18kb
KpnI
SacII
BssHII
EagI
NaeI
NarI
SmaI
BamHI
BamHI + BssHII
BamHI + EagI
BamHI + SacII
BamHI + SmaI
BamHI + NaeI
23kb
9kb
6kb
4kb
Probe A
KpnI
NarI + KpnI
NaeI + KpnI
SmaI + KpnI
SacII + KpnI
EagI + KpnI
BssHII + KpnI
23kb
9kb
6kb
4kb
Probe B

Figure 4.4 Southern blot of methylated BssHII cleavage site

A map of the Surfeit locus covering Surf-2 and Surf-4 is shown above with the BssHII and HindIII sites indicated. Beneath is a Southern blot of mouse cellular DNA cleaved with BssHII alone, HindIII and BssHII or HindIII alone. The position of the probe used is indicated on the map above, and it is the fragment labelled SEB1 Pst in appendix 2a. Each lane contains 10 µg of Balb/C 3T3 A31 DNA and the right hand lane shows the lambda HindIII fragments used as size markers.


Surf-4
Surf-2
CpG
1.0kb
HindIII
BssHII
HindIII
Probe
SacII
BssHII
EagI
NaeI
NarI
SmaI


BssHII
BssHII+ HindIII
HindIII
M
23kb
9.4kb
6.5kb
4.3kb

4.3 Results: Location of the Neighbouring CpG-Rich Islands

Having found that all the defined genes in the Surfeit cluster are associated with CpG-rich islands, and that there are 4 CpG-rich islands within the region, it was interesting to know the extent of this clustering of genes and CpG-rich islands. In order to look for islands outside the Surfeit cluster, unique probes from outside the left-hand and right-hand most islands were used to probe Southern blots of mouse cellular DNA cleaved with island specific rare cutting enzymes. Neither probe detected any fragments on these Southern blots indicating that the next cutting sites for all the enzymes are further than about 40 kb away.

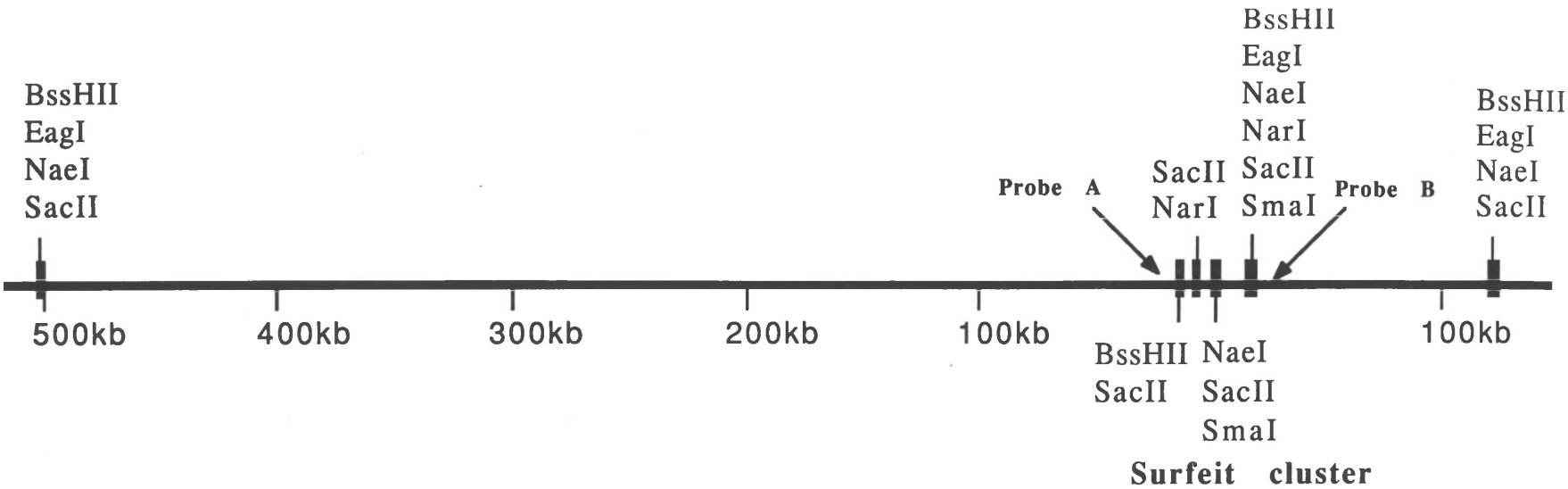
In order to resolve larger fragments pulsed field gel electrophoresis was used to fractionate DNA digested with the island specific rare cutting enzymes. The genomic DNA was prepared from mice spleens rather than from Balb/C 3T3 cells as this is less likely to be undermethylated, and the DNA was prepared in low melting agarose blocks to prevent shearing.

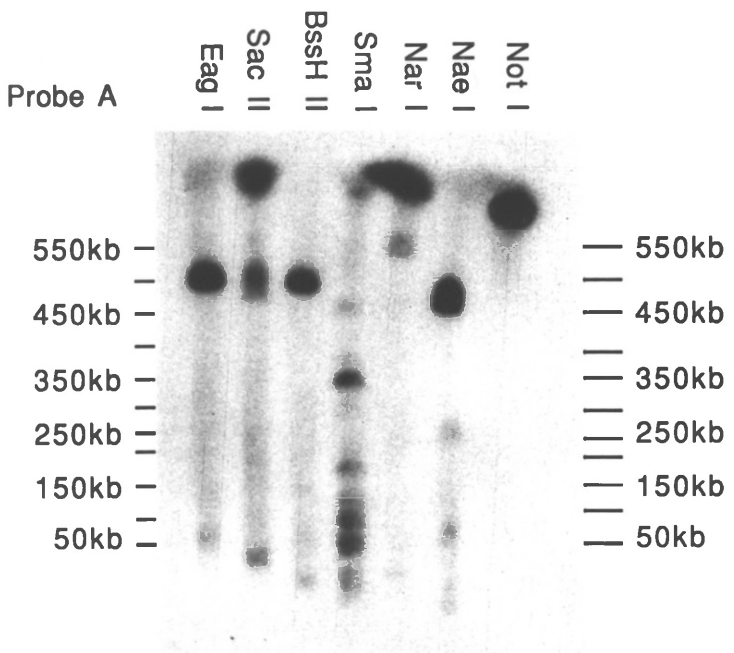
Figure 4.5 shows two such pulsed field gels blotted. The left hand blot is probed with a fragment from the far left hand end of the Surfeit locus (probe A in Fig. 4.5) and the right hand blot is probed with a fragment located at the far right hand side of the Surfeit locus, 6 kb upstream of Surf-4 (probe B in Fig. 4.5). Probe A detects EagI, SacII, BssHII and NaeI fragments of 500 kb indicating that the neighbouring island on the left hand side is 500 kb distant. Probe B detects EagI, SacII, BssHII and NaeI fragments of 100 kb in size indicating that this is the position of the neighbouring island on the right-hand side. The smaller SmaI fragment detected with probe B is probably due to an unmethylated SmaI site not located in a CpG-rich island as only 42% of SmaI sites are expected to occur in islands (Table 1.1). Figure 4.5 also shows a long range map of the relative spacing of the CpG-rich islands located in the vicinity of the Surfeit cluster.

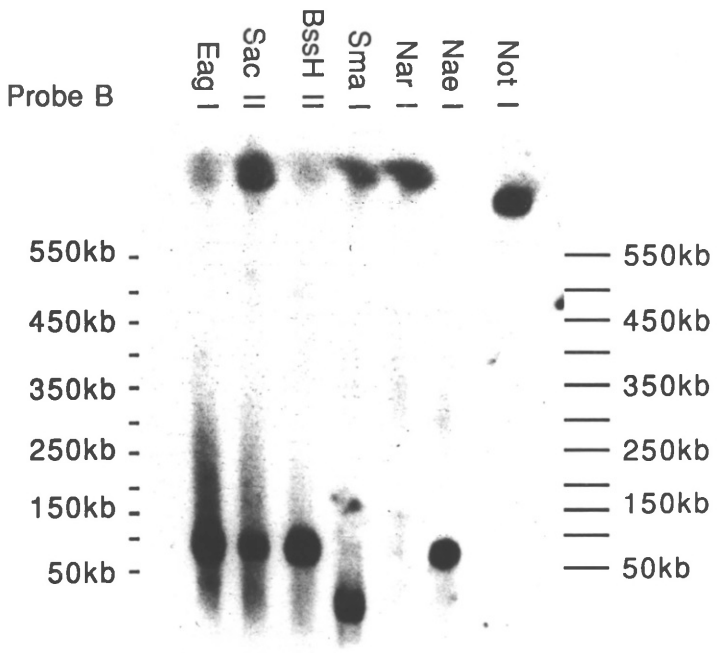
Figure 4.5 CpG-Rich islands neighbouring the Surfeit Cluster

Above is shown a map of the genomic DNA surrounding the Surfeit cluster. Distances are given in kb and CpG rich islands are shown as solid boxes. The island specific rare cutting enzyme sites which cleave in each island are listed above or below the island. The map is centred on the island between Surf-1 and Surf-2.

Below are shown two pulsed field gels blotted and probed with the fragments indicated. The position of Probe A is indicated on the map above and is the Xho1 to HindIII fragment labelled in appendix 3. The position of probe B is shown on the map above and is an 800 bp Apa1 fragment located 6 kb upstream of Surf-4 and labelled in appendix 2a. Each track contains DNA from $2x10^6$ cells from the spleen of a DBA mouse and the gel was run at 170 V with a switching time of 20 secs for 24 hrs. The sizes on the left indicate the positions of concatermerized bacteriophage lambda DNA.


BssHII
EagI
NaeI
SacII
500kb
400kb
300kb
200kb
100kb
Probe A
SacII
NarI
BssHII
EagI
NaeI
NarI
SacII
SmaI
Probe B
BssHII
EagI
NaeI
SacII
100kb
BssHII
SacII
NaeI
SacII
SmaI
Surfeit cluster


Probe A
Eag I
Sac II
BssH II
Sma I
Nar I
Nae I
Not I
550kb
450kb
350kb
250kb
150kb
50kb
550kb
450kb
350kb
250kb
150kb
50kb


Probe B
Eag I
Sac II
BssH II
Sma I
Nar I
Nae I
Not I
550kb
450kb
350kb
250kb
150kb
50kb
550kb
450kb
350kb
250kb
150kb
50kb

4.4 Results: Location of CpG-Rich Islands Neighbouring Ribosomal Protein Genes

The only Surfeit gene to be identified so far is Surf-3 which encodes ribosomal protein L7a. As Surf-3 occurs in a very unusual cluster of genes it is of interest to know whether the other ribosomal protein genes also occur in such clusters and indeed whether clusters are a feature of the control of ribosomal protein (rp) genes. As described in chapter 2.4, five other mammalian, intron containing ribosomal protein genes have been cloned. In the case of the rpS16 and rpL30 genes a unique DNA fragment from either side of each gene was used to probe Northern blots but no mRNAs were detected (Wagner & Perry 1985, Wiedemann & Perry 1984). Although this might indicate that these two genes are not located in clusters of genes, the results could also be due to the probes being derived from introns or to the neighbouring genes not being expressed in the cell type from which the mRNA, that was used for the Northern blots, was extracted.

As the 5' end of each published ribosomal protein gene is located in a CpG-rich island I decided to investigate whether there are other CpG rich islands nearby, using the same type of analysis which I used to locate islands in the Surfeit locus. Unique probes from either side of the rpL30 and rpL32 genes were made using the Polymerase Chain Reaction (PCR) to amplify pieces of DNA from the published sequences. These pieces were subcloned before using to probe Southern blots, because, if the PCR amplified DNA was labelled directly from a Low Melting Agarose gel they gave a high background due to the presence of repetitive DNA. When subcloned and used as probes, these pieces detected unique fragments on Southern blots. They also hybridized to fragments of the expected size for the intron-containing rpL30 and rpl32 genes on a Southern blot of mouse DNA digested with KpnI, XhoI, XbaI, SacI, HindIII, BamHI or EcoRI, indicating that the correct pieces of DNA had indeed been cloned.

For rpL30 the sequence published by Wiedemann & Perry (1984), shows a SacII and an EagI site in the CpG-rich island. DNA from nt 181 to 461 of the published sequence, located 3' to the island, was subcloned and used to probe a Southern blot of double digests of BamHI and island specific rare cutting enzymes. This showed that the island contains a SacII and a SmaI site rather than an EagI site. This could be due to a polymorphism between different mouse strains or due to a mistake in the published sequence, mistakes being easily made where runs of C and G cause compactions on the sequencing gel. When used to probe a Southern blot of mouse DNA digested with island specific rare cutting enzymes, which resolves DNA fragments up to about 40 kb in size, no bands were detected indicating that the fragments generated by the enzymes are larger than about 40 kb. When used to probe a pulsed field gel of mouse DNA digested with island specific rare cutting enzymes, very large BssHII, NaeI and EagI fragments, a 250 kb SacII fragment, and 50 kb and 250 kb SmaI fragments were detected (Fig. 4.6).

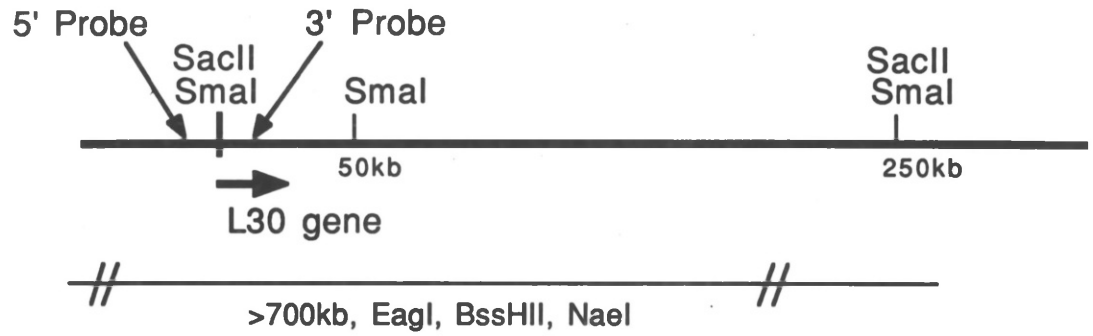
A second probe located 5' to the island at the 5' end of the rpL30 gene was also made running from nt -438 to -107 of the published sequence. This probe was shown to be unique and again detected a SacII and SmaI site in the island at the 5' end of the rpL30 gene on a blot of double digests of BamHI and rare cutting enzymes No fragments smaller than about 40 kb were detected with cellular DNA cut with any of the island specific rare cutting enzyme separated by normal gel electrophoresis. However, on a pulse field gel, the same large BssHII, EagI and NaeI fragments detected with the 3' probe were again detected with the 5' probe (Fig. 4.6). No fragments were detected for SacII, SmaI, NarI and NotI on this pulsed field gel and this could be due to the fragments being too large to be resolved. Figure 4.6 shows a long range map of island specific rare enzyme cutting sites around the rpL30 gene and two pulsed field gels probed with the 5' and 3' probes. Clearly the mouse rpL30

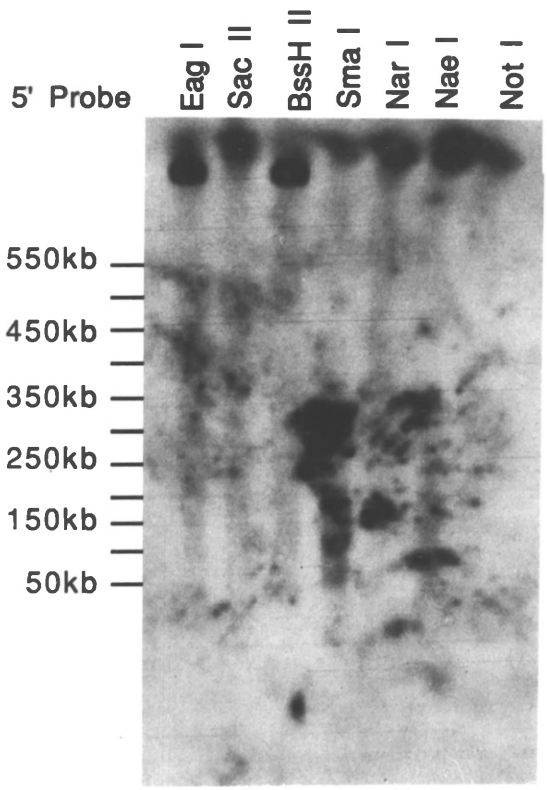
gene does not have closely neighbouring CpG-rich islands and does not lie in a cluster of genes and CpG-rich islands resembling the Surfeit cluster.

For the rpL32 gene the sequence published by Dudov & Perry (1984) contains BssHII, SacII and SmaI sites in the CpG-rich island. A piece of DNA located 5' to the island from nt -350 to -39 in the published sequence was amplified by PCR and subcloned. On a Southern blot of double digests of BamHI and island specific rare cutting enzymes, the BssHII, SacII and SmaI sites in the CpG-rich island were confirmed. On a Southern blot which resolved fragments up to about 40 kb, no fragments were detected with the rare cutting enzymes indicating that the fragments generated with these enzymes are larger than about 40 kb. On a pulsed field gel of cellular DNA digested with island specific rare cutting enzymes, 300 kb and 400 kb EagI, 325 kb and 400 kb SacII, 325 kb BssHII, 200 kb and 325 kb NaeI fragments were detected. Figure 4.7 shows this blot and a long range map around the rpL32 gene, it is clear that rpL32 has no close islands on the upstream side of the gene.

Figure 4.6 Long range map around the ribosomal protein L30 gene.

Above is shown a map of the island specific rare cutting enzyme sites around the mouse ribosomal protein L30 gene. Below is a pulsed field gel with mouse DNA digested with EagI, SacII, BssHII, SmaI, NarI, NaeI, and NotI. On the left the blot has been probed with the 5' probe and on the right the blot has been probed with the 3' probe. The 5' probe runs from -438 to -107 and the 3' probe runs from positions 181 to 461 of the sequence published by Wiedemann and Perry (1984). Each lane contains DNA from $2x10^6$ cells from the spleen of a DBA mouse, the size markers were concatermerized phage lambda DNA.


5' Probe
3' Probe
SacII
SmaI
SmaI
SacII
SmaI
50kb
250kb
L30 gene
>700kb, EagI, BssHII, NaeI


5' Probe
Eag I
Sac II
BssH II
Sma I
Nar I
Nae I
Not I
550kb
450kb
350kb
250kb
150kb
50kb

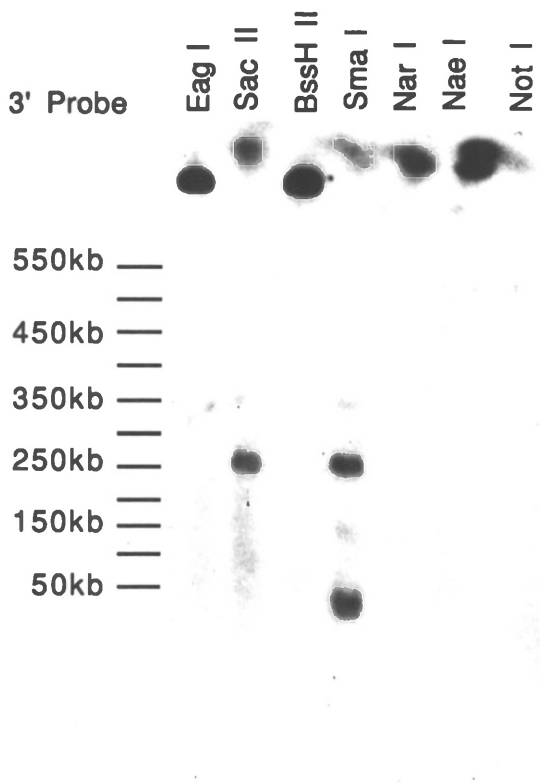
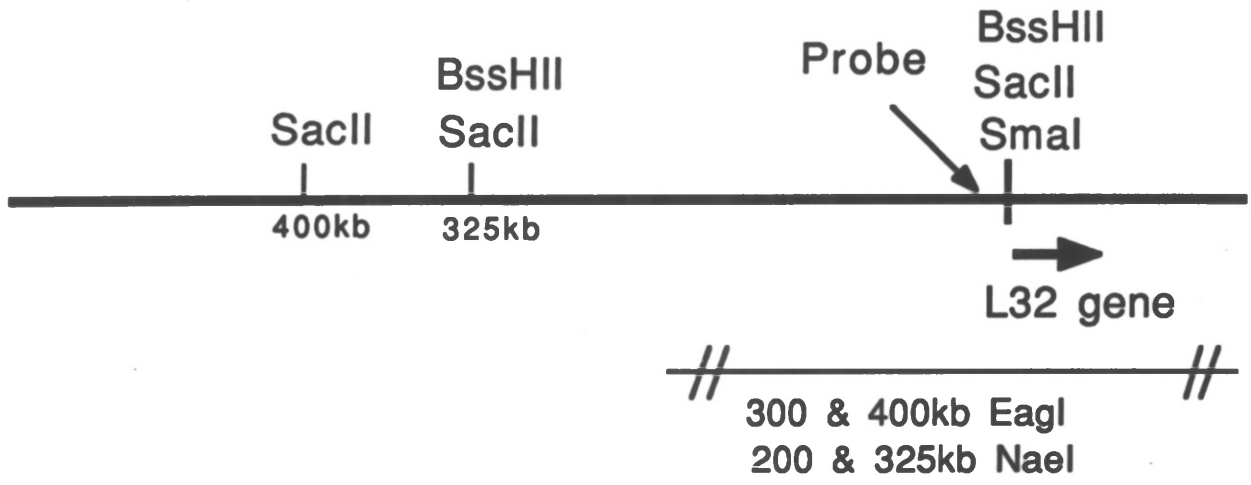
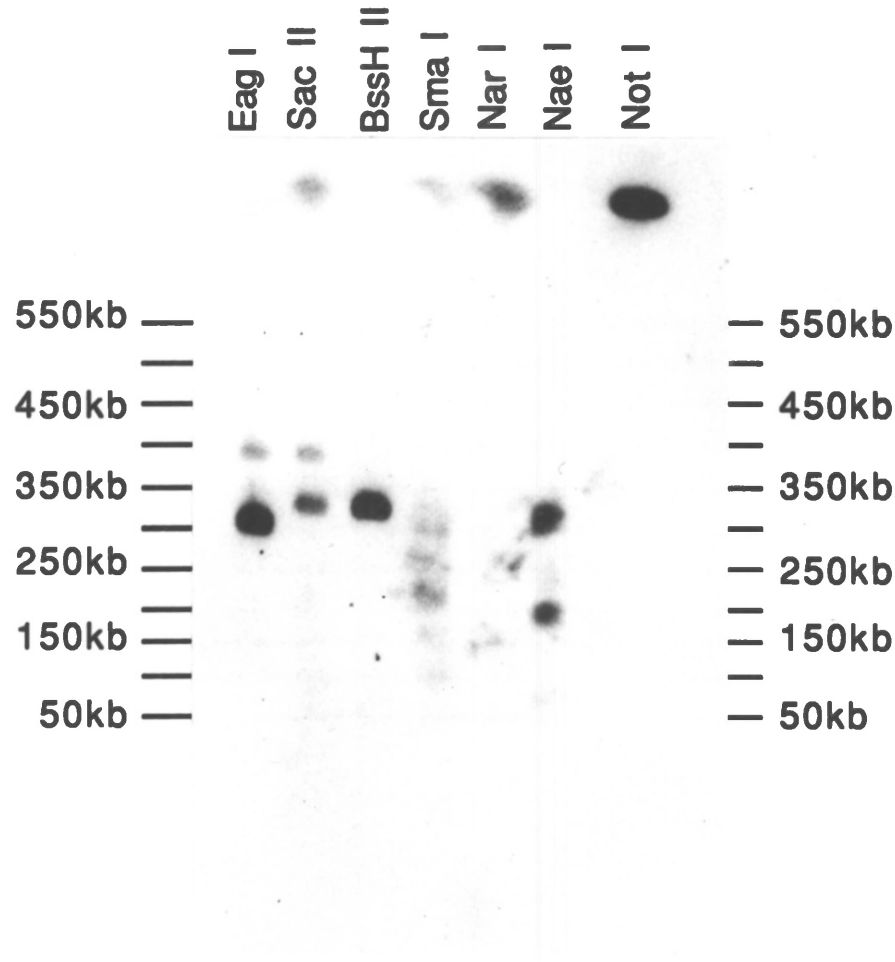

3' Probe
Eag I
Sac II
BssH II
Sma I
Nar I
Nae I
Not I
550kb
450kb
350kb
250kb
150kb
50kb

Figure 4.7 Long range map around the ribosomal protein L32 gene

Above is shown a map of the island specific rare enzyme cutting sites around the mouse ribosomal protein L32 gene. Below is a pulsed field gel with mouse DNA digested with EagI, SacII, BssHII, SmaI, NarI, NaeI, and NotI. The blot has been probed with a DNA fragment running from position -350 to -39 in the sequence published by Dudov and Perry (1984). Each lane contains DNA from $2x10^6$ cells from the spleen of a DBA mouse, the size markers were concatermerized phage lambda DNA.


BssHII
SacII
SmaI
Probe
SacII
BssHII
SacII
400kb
325kb
L32 gene
300 & 400kb EagI
200 & 325kb NaeI


Eag I
Sac II
BssH II
Sma I
Nar I
Nae I
Not I
550kb
450kb
350kb
250kb
150kb
50kb
550kb
450kb
350kb
250kb
150kb
50kb

4.5 Discussion

Four CpG-rich islands have been identified in the Surfeit cluster, one at the 5' end of each identified gene, and one 8 kb to the left of Surf-3. Four islands within 32 kb of genomic DNA is very unusual, as described in the introduction (section 1.2). The average spacing of islands is estimated, and indeed observed in long distance mapping (Brown & Bird 1986), to be 100 kb. The presence of an island 8 kb to the left of Surf-3 strongly suggests the presence of one or more genes situated on this side of the Surfeit cluster which it is clearly of interest to identify. There may also be a gene upstream of Surf-4 if this CpG-rich island contains a bidirectional promoter. The locations of the neighbouring islands 100 kb and 500 kb on either side of the Surfeit cluster indicates that we may have come to the end of the very tightly clustered region. However, it does not rule out the presence of genes which lack CpG-rich islands nor the presence of very extended genes which could span the distance between the Surfeit locus and the neighbouring islands.

Probes made from the published sequences of ribosomal proteins L32 and L30 that were used to probe pulsed field gel blots of genomic DNA cut with island specific rare cutting enzymes, indicate that these genes are not located in clusters of CpG-rich islands similar to the Surfeit cluster. This means that although Surf-3 is a ribosomal protein gene located in a cluster of CpG-rich islands and genes, clustering does not appear to be a common feature of ribosomal protein genes.

Chapter 5: Identification of Two New Genes and Repetitive Elements in the Surfeit Cluster

5.1 Introduction

Chapter 4 showed that the 5' end of each of the identified Surfeit genes is associated with a CpG-rich island and that there is one additional island of about 1.0 kb situated 7 kb upstream of Surf-3. Investigations were undertaken to determine whether the 8 kb of DNA upstream of Surf-3 is transcribed and whether the CpG-rich island is associated with the 5' end of a gene. The 8 kb of DNA upstream of Surf-4 which has also been cloned could also contain other Surfeit genes.

5.2 Results: Two New Genes

The 8 kb of cloned genomic DNA upstream of Surf-3 was mapped. Various fragments derived from it (Fig. 5.1), were used to probe Southern blots of mouse genomic DNA to determine whether they are unique. The whole region had been previously probed with total genomic DNA (Williams 1986) but this technique does not distinguish between low copy and unique genomic DNA, and some fragments which had appeared unique are in fact repetitive. Those fragments which were found to be unique were used to probe Northern blots of 5 µg of Poly(A)+ and 20 µg of Poly(A)- RNA from DBA L1210 cells to determine whether any transcripts present in this cell line are derived from this region of the Surfeit locus. Although this technique would detect transcripts expressed in leukemic cells, there may be other transcripts expressed in other differentiated cell types which would not have been detected.

Three transcripts of 1.0, 2.4 and 3.0 kb in size were detected with the various unique probes. Probes 7, 8, 10 and 11 all detected a 3.0 kb poly(A)+ RNA. Probe 11 also detected a 1.0 kb mRNA which could be a shorter spliced

product of the gene specifying the 3.0 kb mRNA. Figure 5.2 shows Northern blots hybridized with probe 8 and probe 11. The transcription unit which has been found to lie in the five kb directly upstream of Surf-3 is now called Surf-5.

Probe 2 does not detect any transcripts at all, but probes 3, 4 and 5 all detect a 2.4 kb poly(A)+ RNA. This 2.4 kb transcript is now called Surf-6 and figure 5.3 shows a Northern blot obtained with probe 5. To determine the direction of transcription of the Surf-6 message, two plasmids were constructed by inserting fragment 4 into the vector pSP65 in both orientations. The vector pSP65 has an SP6 RNA polymerase promoter upstream of the polylinker used for insertion of the fragment, and this promoter was used to generate single stranded RNA labelled with ^{32}P complementary to both strands of the fragment 4 DNA. This labelled RNA was used to probe a northern blot and only RNA generated from one of the two plasmids was found to hybridize to the 2.4 kb mRNA (Fig. 5.3). The orientation of fragment 4 in the plasmid was such that it would hybridize to a mRNA transcribed in the opposite orientation to Surf-3 (see Fig. 5.3). As this 2.4 kb mRNA is not detected with probes closer to Surf-3 than fragment 5 the 5' end of Surf-6 must be close to the CpG-rich island, and is most likely associated with it.

It is of interest to note that one cDNA generated from RNA extracted from COS cells transfected with the 19 kb EcoRI fragment, contained a 5' end homologous to 200 bp in this CpG-rich island. The cDNA, called C3B (see appendix 1b for a map), has 200 bp at its extreme 5' end which spans the BglII site in the CpG-rich island upstream of Surf-3 and is transcribed away from Surf-3. The 5' end of C3B is spliced onto the first exon of Surf-3 suggesting that it was derived from a rearranged molecule of transfected DNA. This is consistent with there being a promoter directing transcription away from this upstream CpG-island.

A similar analysis was carried out on the 8 kb of DNA upstream of Surf-4 to determine which fragments of DNA in this region are unique. Most of the sequences of this 8 kb are repetitive with only a single Apa1 fragment of 800 bp being found to be unique (see appendix 2a).

Figure 5.1 Map of the 8 kb of DNA upstream of Surf-3.

The central bar shows the cloned genomic DNA, open bar indicates repetative DNA and solid bar indicates unique DNA. Distances are marked in kb and the position of selected restriction enzyme recognition sites are shown. Below are shown the fragments used as probes, each identified by a number (1-11). Above are indicated the Surf-3, Surf-5 and Surf-6 transcription units, with the direction of transcription of Surf-3 and Surf-6 indicated. The two CpG-rich islands are shown as solid boxes. The position of the repetitive B2 element discussed in section 5.3 is indicated.

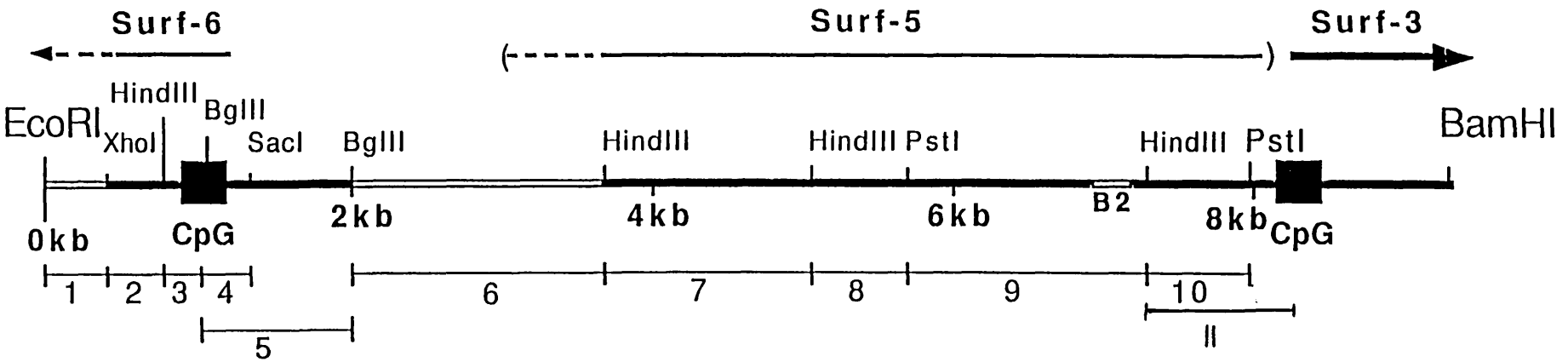


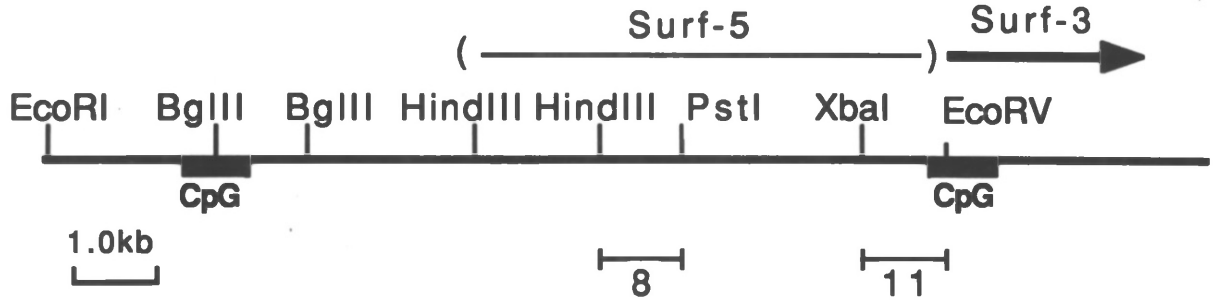
Figure 5.2 The Surf-5 transcription unit.

Above is shown a map of the region upstream of Surf-3 with the position and extent of the Surf-5 transcription unit indicated. The location of probes 8 and 11 is shown beneath the map.

a) Northern blot hybridized with probe 8 used as a double stranded DNA probe.

b) Northern blot hybridized with probe 11 used as a double stranded DNA probe.

In both blots, the left lane contains 5 μg of poly(A)+ RNA and the right lane contains 20 μg of poly(A)- RNA from DBA L1210 cells. The sizes on the left indicate the positions of RNA size markers (BRL).


Surf-5
Surf-3
EcoRI
BglII
BglII
HindIII
HindIII
PstI
XbaI
EcoRV
CpG
CpG
1.0kb
8
11

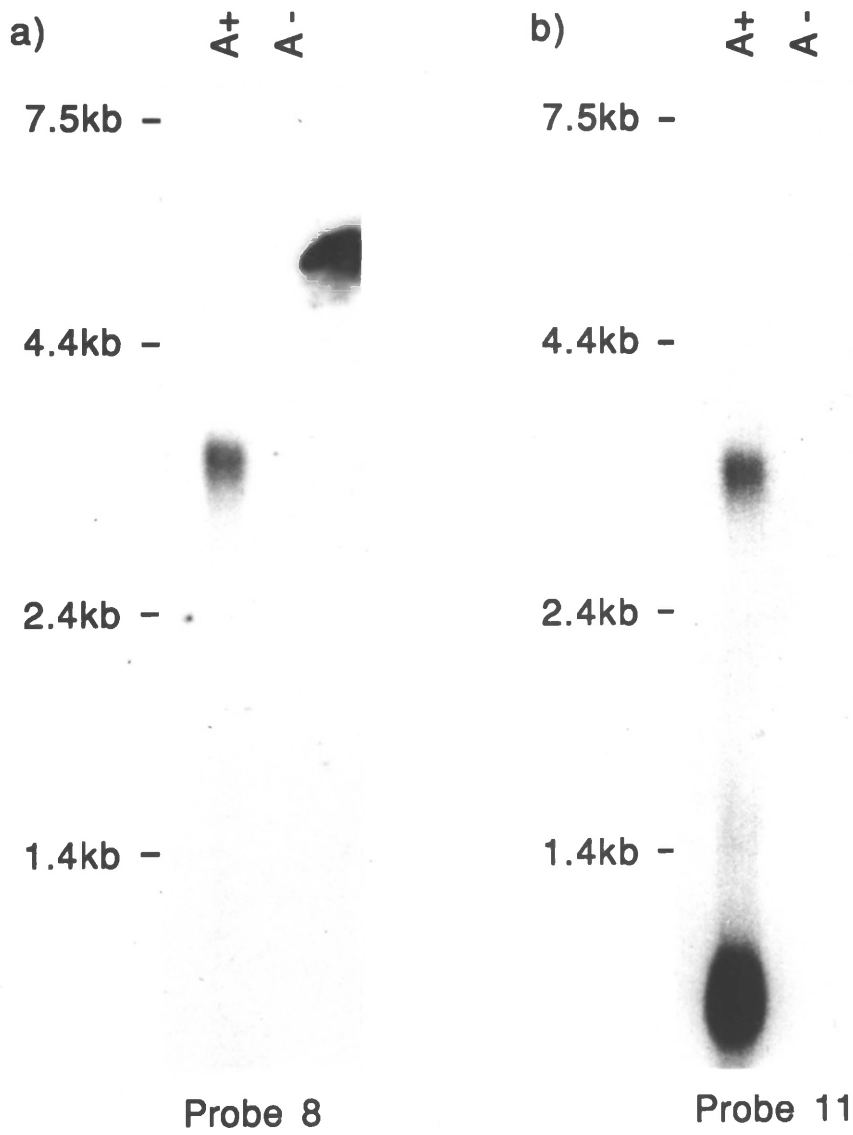

a)
A+
A-
7.5kb -
4.4kb -
2.4kb -
1.4kb -
Probe 8
b)
A+
A-
7.5kb -
4.4kb -
2.4kb -
1.4kb -
Probe 11

Figure 5.3 The Surf-6 transcription unit.

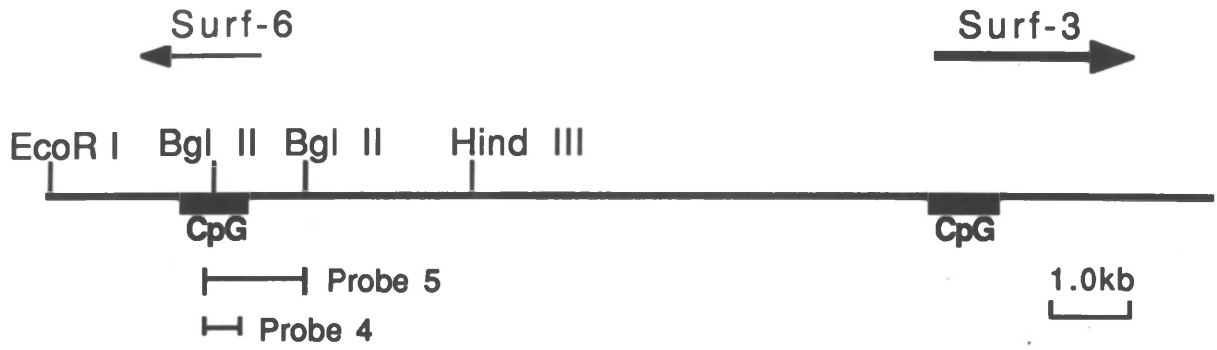
Above is shown a map of the region upstream of Surf-3 with the position and extent of the Surf-6 transcription unit. The locations of probes 4 and 5 are indicated beneath the map.

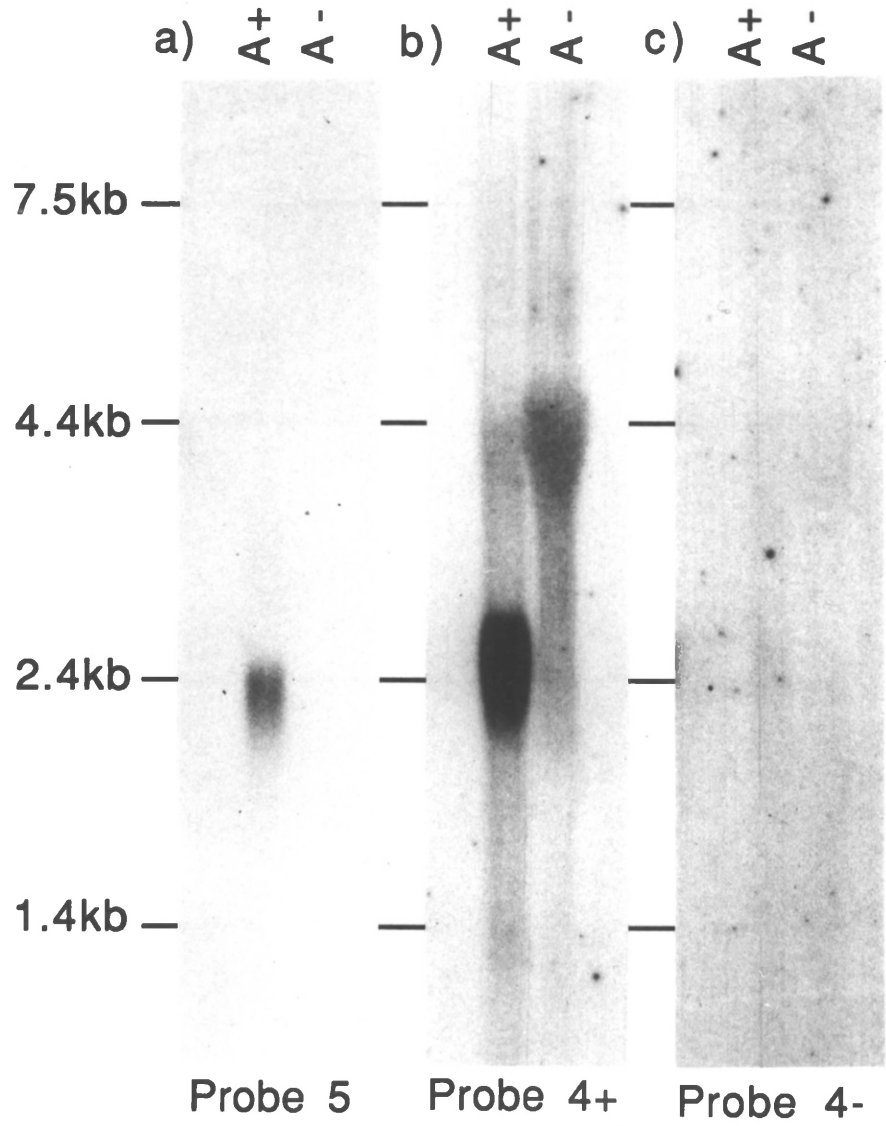
a) Northern blot loaded with 5 µg Poly(A)+ or 20 µg Poly(A)- RNA from DBA L1210 cells, hybridized with probe 5 as a double stranded DNA probe.

b) Northern blot as above probed with a single stranded RNA probe derived from fragment 4.

c) Northern blot as above probed with a single stranded RNA probe derived from the other strand of fragment 4 used in b).

For all the Northern blots the sizes on the left indicate the positions of RNA size markers (BRL).


Surf-6
Surf-3
EcoR I
Bgl II
Bgl II
Hind III
CpG
CpG
Probe 5
Probe 4
1.0kb


a)
A+
A-
b)
A+
A-
c)
A+
A-
7.5kb
4.4kb
2.4kb
1.4kb
Probe 5
Probe 4+
Probe 4-

5.3 Results: Repetitive Elements

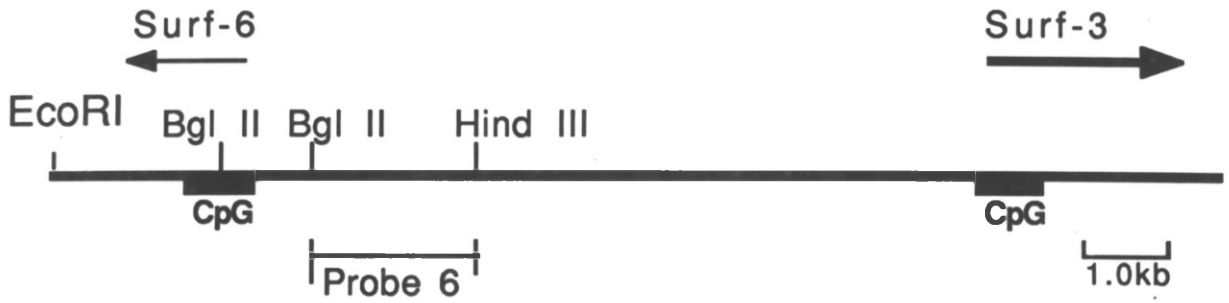
Fragment number 6 in figure 5.1 is nearly unique, but hybridizes to a number of bands in addition to the predominant Surfeit locus band on a Southern blot of Balb/C DNA digested with a number of enzymes (Fig. 5.4). This fragment did not appear repetitive when probed with total genomic DNA and it probably contains a sequence that is repeated in only a few other places in the mouse genome. A Northern blot probed with fragment 6 shows a mRNA of 3.0 kb which probably corresponds to Surf-5 though it could be, all or in part, a 3.0 kb mRNA derived from another region of the mouse genome (Fig 5.4).

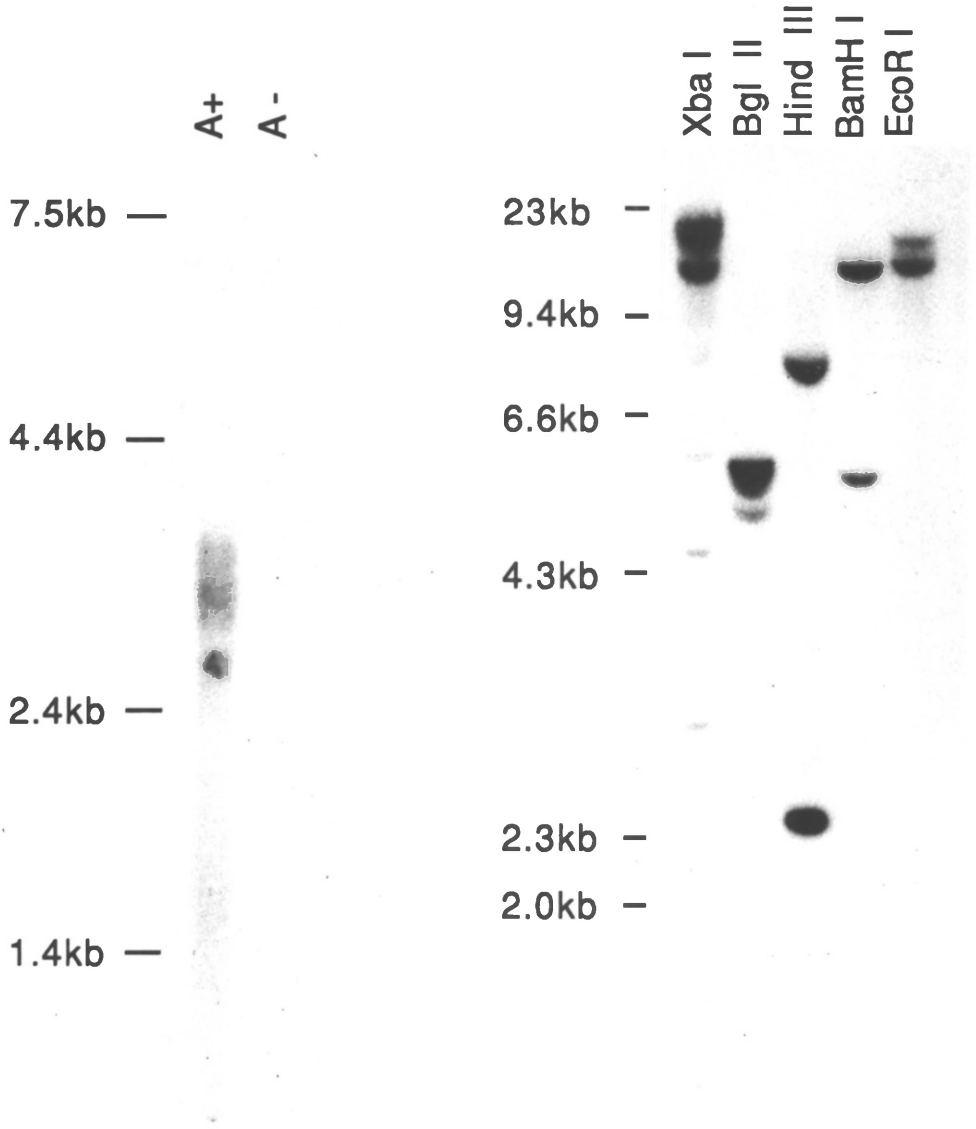
Figure 5.4 Low level repetitive DNA upstream of Surf-3.

Above is shown a map of the region upstream of Surf-3 with fragment 6 indicated as the probe.

Underneath, on the left, is a Northern blot with 5 µg of poly(A)+ and 20 µg poly(A)- RNA from DBA L1210 cells, probed with fragment 6. The size markers were RNA markers (BRL).

On the right is a Southern blot of Balb/C A31 DNA (5 µg/lane) cut with the enzymes indicated and probed with fragment 6. The numbers on the left indicate the positions of HindIII fragments of Lambda phage DNA.

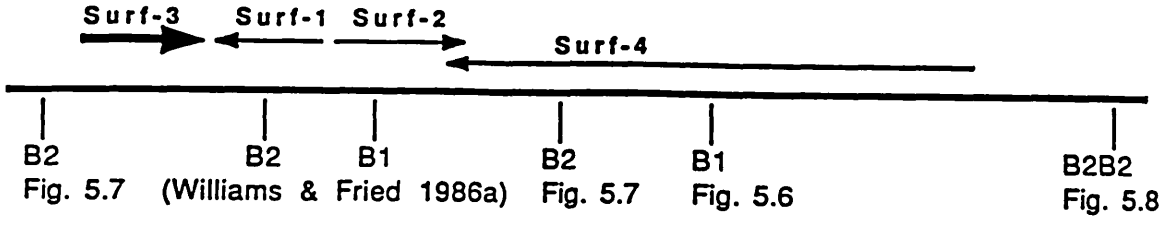

Surf-6
Surf-3
EcoRI
Bgl II
Bgl II
Hind III
CpG
CpG
Probe 6
1.0kb


A+
A-
7.5kb
4.4kb
2.4kb
1.4kb
Xba I
Bgl II
Hind III
BamH I
EcoR I
23kb
9.4kb
6.6kb
4.3kb
2.3kb
2.0kb

Repetitive elements were initially detected by probing southern blots of cloned fragments with labelled total mouse genomic DNA, only those fragments which are very highly repetitive hybridize to give a detectable radioactive signal. Several of these repetitive elements were sequenced and found to be either B1 or B2 elements. B1 and B2 elements are the most abundant short interspersed sequences (SINES) found in Rodent DNA, each occurring in about 100,000 copies (Rogers 1985) (see section 1.2). Figure 5.5 shows a map of the whole Surfeit cluster with the approximate positions of the identified B1 and B2 elements.

Figure 5.5 B1 and B2 elements in the Surfeit gene cluster.

The continuous line represents the mouse DNA containing the Surfeit cluster with arrows indicating the four Surfeit genes above. The positions of the six B1 and B2 elements are indicated. The two repetitive elements previously published (Williams & Fried 1986a) are indicated. Underneath each element discussed in this chapter, is the figure number in which the sequence is given. The positions of the four elements discussed in this chapter are given in more detail in appendix 2a and 3.



B1 elements are related in sequence to the Alu sequences in human DNA and are thought to have arisen from the 7SL RNA (part of the signal recognition particle which acts in protein secretion) (Ulla and Tschudi 1984). The consensus sequence for B1 elements in mouse (Kalb et al 1983) is shown in figure 5.6 and is aligned with pieces of a human 7SL cDNA (Ullu & Tschudi

1984) and with a B1 element located in the 2nd intron of Surf-4 (the position of this B1 element is shown in appendix 2a). At the 3' end of the B1 element there is a variable A rich region which is thought to be important for the priming of reverse transcription by hybridization between this A rich region and the U rich termination region of the RNA molecule (Jagadeeswaran et al 1981). The direct repeats, which normally flank B1 elements (Rogers 1985), are discernible as imperfect repeats of 20 bp on either side of the element and are underlined. This B1 elements is 12% diverged from the consensus B1 sequence and this is typical of B1 elements which differ on average 13% from the consensus sequence (Rogers 1985)

B2 elements are thought to have arisen from a tRNA, possibly a serine tRNA (Daniels and Deininger 1985). Figure 5.7 shows the consensus sequences for the two subfamilies of mouse B2 elements (Rogers 1985), these are aligned with a Rat serine tRNA (Daniels and Deininger 1985) and two B2 elements from the Surfeit cluster. One is a partial element located just upstream of exon 6 of Surf-4 (see appendix 2a) which is truncated at the 3' end. The other is located about 1 kb upstream of Surf-3, and has only been partially sequenced (see appendix 3), the 5' end of this element is embedded in the A-rich sequence characteristic of the 3' end of either a B1 or a B2 element oriented in the other direction. This indicates that this B2 element has inserted into the 3' end of an element already present which is commonly observed (Rogers 1985). Both B2 elements belong to subfamily I identified by Rogers (1985) and as neither element is complete at both ends the flanking direct repeats cannot be seen. The partial B2 element is highly diverged being 23% different to the consensus B2 sequence, whereas the element upstream of Surf-3 is 12% different. On average mouse B2 elements are 8.8% different from the consensus sequence.

Figure 5.8 shows the sequence upstream of Surf-4 where there are two adjacent B2 elements (see appendix 2a). It is quite frequently observed that

B1 and B2 elements integrate into other elements which are already present (Rogers 1985) and this is what appears to have happened here. The sequence does not extend far enough to determine whether this pattern of adjacent repetitive elements continues but it could explain why so much of the DNA upstream of Surf-4 has been found to be repetitive. Where one element has integrated into another one, it is often not possible to detect the direct repeats which are normally found at both ends of B2 elements (Rogers 1985). Both elements belong to subfamily I and are 11 and 14% different to the consensus B2 sequence.

Figure 5.6 B1 Element in the Surfeit cluster.

The consensus sequence of the B1 repetitive element of the mouse (Kalb et al 1983) aligned with stretches of a human 7SL cDNA (Ullu & Tschudi 1984), and a B1 element located in the second intron of Surf-4 (see appendix 2a). The imperfect repeats of 20bp which flank the B1 element are underlined. - indicates the same nucleotide as is present in the consensus, ^ indicates a gap, extra base pairs are indicated in the line above or below, and XX indicates an indefinite number of nucleotides

7SL Part of the human 7SL RNA sequence.
con mouse B1 consensus sequence.
Surf sequence of B1 element found in the second intron of Surf-4.

```
7SL                                                          G-------C
con                                                           CCGGGCGT
Surf                                GCAAGGGCAAGAAAGCTGTCACCAGGTT------
                                                                     |
                                                                     G

                                        T                          T
                                        |                          |
7SL  --------GT----G--G-------|----------T-----T--A----CG-|-----C
con  GGTGGCGCACGCCTTTAATCCCAGCACTCGGGAGGCAGAGGCAGGCGGATTTCTGAGTT
Surf ---------T-------------------C---------------^TA-----------

          XXX                       XXXXXXXXX
          |                         |
7SL  -^---|---------G-A---T--C---|          |----T----A---T--C--G---
con  CGAGGCCAGCCTGGTCTACAGAGTGAGTTCCAGGACAGCCAGGGCTACACAGAGAAACC
Surf ------T-------------^--------^--------G---------T---------T-
                                             |
                                             G

7SL  -C----CT
con  CTGTCT
Surf -----GAAAAACAAAAACAAAAACAAAAAAAGAAAAAAAAGAAAGCAAGCAAGAAAACC

Surf AATGTGAGGCCCCAAATAT
```

Figure 5.7 Two B2 elements in the Surfeit cluster.

The consensus sequence for subfamily I of the mouse B2 element is given in full (Rogers 1985). Nucleotides which are different in subfamily II are indicated in the line below (Rogers 1985). Aligned with the B2 consensus sequence is the sequence of the rat serine tRNA gene (Daniels & Deininger 1985), and two B2 elements from the Surfeit cluster. The partial B2 element is located just upstream of the sixth exon of Surf-4 as indicated in appendix 2a and the other, which is only half sequenced, is located about 1 kb upstream of Surf-3 as indicated in figure 5.1 and appendix 3. - indicates identity with the consensus for subfamily I, ^ indicates a gap in the sequence and extra bases are indicated in the line below.

tRNA	sequence of a rat serine tRNA gene.
con I	B2 consensus sequence, subfamily I.
con II	B2 consensus sequence, subfamily II.
par	sequence of a partial B2 element just upstream of exon 6 of Surf-4.
S3	sequence of B2 element upstream of Surf-3.

```
tRNA
con I                                                              GGGGC
conII
par                                                      AAGAAGTATGGCTGCAT
S3                              TTTTCTTTGCTTTTTAAGATTTATTTATTTAAGGA-


tRNA    -AC---G----GG----------GCG-^-G-------AA----TT-TGCTCTGCACG
con I TGGAGAGATGGCTCAGTGGTTAAGAGCACT^GACTGCTCTTCCAAAGGT
conII    T               G          CC              G
par    -------------TG---------------^-----T----T--G---A
                                                    T
S3     ------A--------^^^--------T---^------C----T-G----


tRNA   -G--G----G-------T-CT-GT-G
con I CCTGAGTTCAAATCCCAGCAACCACATGGTGGCTCACAACCATCTGTAATGAGATCTG
conII   A                                                C    C
par    --CAG----------A----C--------CA-------^^^^^------  -------
                                                        CCTCCA
S3     ------C----T-----A-----------------------------------------


con I ATGCCCTCTTCTGGTGTGTCTGAAGACAGCTACAGTGTACTTATATATAATAAATAAA
conII  CT           A                                  C
par    -CTGACACCCTCACATATACATAC


con I TAAA(A)0-8TC(T)2-5(A)7-14
```

Figure 5.8 Adjacent B2 elements upstream of Surf-4.

The B2 subfamily I consensus sequence (Rogers 1985) is written out in full and then repeated with the 5' end truncated. Aligned with the two B2 consensus sequences, is the sequence of two adjacent B2 elements located upstream of Surf-4 (see appendix 2a). - indicates identity with the consensus sequence, ^ indicates a gap in the sequence and extra bases are in the line below.

```
con       consensus sequence for B2 elements
2nd       consensus sequence for B2 elements repeated
Surf      sequence of two contiguous B2 elements upstream of Surf-4

con                                                        GGGGCTGGA
Surf                                  CAACCTGTCTCAAAAACCAAACA-------G

con  GAGATGGCTCAGTGGTTAAGAGCACTGACTGCTCTTCCAAAGGTCCTGAGTTCAATCCC
surf ---------T--CA-----------------------C---------------A-TT

con  AGCAACCACATGGTGGCTCACAACCATCTGTAATGAGATCTGATGCCCTCTTCTGGTCT
surf ------^--------------------CA-------------C-------------G-
     |
     T

con  GTCTGAAGACAGCTACAGTGTACTTATATATAATAAATAAATAAATC(T)2-5(A)7-14
Surf ----------------G---------GTC-------^--------TCTTTAGGC

2nd  GCTCTTCCAAAGGTCCTGAGTTCAAATCCCAGCAACCACATGGTGGCTCACAACCATCT
Surf CGGAGTAAGC----------A-------T----------A--CA--------------^^
                  |              |
                  G              T

2nd  GTAATGAGATCTGATGCCCTCTTCTGGTCTGTCTGAAGACAGCTACAGTGTACTTATAT
Surf ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^N---------T--------C----
                                            |
                                            T

2nd  ATAATAAATAAATAAA(A)0-8  TC(T)2-5(A)7-14
Surf -C-T------------TAAATAAATCTTAATAAAACAAACAAACTTAAAGCACTTATTCT

Surf AAAATAACATAAAGAATAAATTGGCCAGTGGTGGTGGT
```

5.4 Discussion and Further Work

A detailed analysis of the 8 kb of cloned DNA upstream of Surf-3 has revealed two new genes in the Surfeit cluster. Surf-5 is located within the 5 kb of DNA directly upstream of Surf-3 and specifies a mRNA of 3.0 kb and a smaller mRNA of 1.0 kb. Surf-6 is located 7 kb upstream of Surf-3 and specifies a mRNA of 2.4 kb. Surf-6 is transcribed from the CpG-rich island away from Surf-3. This brings the total number of Surfeit genes to six and extends the region of clustered genes and CpG-rich islands.

A B1 and four B2 elements have been identified in the Surfeit cluster in addition to the B1 element in Surf-2 and the B2 element in Surf-1 which were previously identified (Williams and Fried 1986a). Two of these B2 elements are adjacent suggesting that the second element integrated into the first. These elements conform well to the consensus sequences and are spread throughout the Surfeit locus (Fig. 5.5). There are various other repetitive fragments of DNA which are not as highly reiterated as the B1 and B2 elements.

It will be interesting to fully characterize the Surf-5 and Surf-6 genes, to sequence their cDNAs and determine their organization on the genomic DNA. Sequencing these genes will tell us whether they are related by DNA sequence or protein sequence identity to the other Surfeit genes or to other genes in the EMBL, Genbank or PIR databases. Determining the differential expression of Surf-5 and Surf-6 will tell us how their regulation is related to the organization of the Surfeit gene cluster.

Determining whether there is a gene upstream of Surf-4 by probing a Northern blot with the unique ApaI fragment will extend our knowledge of the organization of the gene cluster.

Chapter 6: Cell Cycle Regulation of the Surfeit Genes

6.1 Introduction

The bidirectional promoter between Surf-1 and Surf-2 may operate either such that both genes are up-regulated together or such that when one gene is up-regulated the other is down-regulated. Similarly the overlap at the 3' ends of Surf-2 and Surf-4 may give antisense regulation of these genes such that when one is on, the other is down-regulated. There are many genes which need to be either co-regulated or inversely regulated with relation to each other, but ones which particularly need temporal regulation with relation to each other are those involved in the cell cycle. To investigate the possibility that the Surfeit genes are differentially regulated through the cell cycle, the steady state levels of the mRNAs of the different Surfeit genes were measured in cells which had been sorted into different stages of the cell cycle. If the layout of the Surfeit genes is related to the regulation of expression of the Surfeit genes through the cell cycle then one would expect to observe differences in the steady state level of the various mRNAs through the cell cycle. However if changes are not observed, this does not rule out that cell cycle regulation of expression of the Surfeit genes is taking place post transcriptionally.

There are four widely used methods of getting populations of cells enriched for the different stages of the cell cycle; mitotic selection, blocking the cells and subsequently releasing the block so the cells progress through the cell cycle, FACS sorting and centrifugal elutriation (Grdina et al 1987). Mitotic selection (Terasima & Tolmach 1961) involves growing certain cell lines in culture, gently shaking the culture vessel dislodges cells which are undergoing mitosis. These cells are collected giving one a synchronous population. This simple method can be applied to a variety of cultured cell lines and gives highly synchronized populations of cells which have

undergone the minimum of manipulation. However, the yield of cells is extremely low, about 2-8% of the population, making it unsuitable for studies of mRNA levels.

Cells can be blocked at various stages of the cell cycle with a wide variety of cytostatic or cytotoxic agents, or by deprivation of essential nutrients, one of the most widely used agents being hydroxyurea (Meyn et al 1975). All the methods of blocking cells suffer from the same major drawback which is the possibility of inducing perturbations in the biochemical balance of the cells giving rise to changes in gene expression which are not related to progression through the cell cycle.

To collect a population of synchronized cells with a Fluorescence Activated Cell Sorter it is necessary to stain the cells with Hoechst dye which stains DNA (Arndt-Jovin and Jovin 1977). The FACS is then set to collect cells with the G1 amount of DNA. A highly synchronized population of cells is achieved, but it is difficult to accumulate large numbers of cells, and the cells have undergone a lot of manipulation, including staining.

Centrifugal elutriation (Meistrich et al 1977) separates cells by their speed of sedimentation. Sedimentation rate is proportional to the two thirds power of the cells volume, as well as to its shape and density. As the volume of a cell increases steadily in an exponential manner with age during the cell cycle, centrifugal elutriation separates a population of cells into fractions, each with a different mean size and stage of the cell cycle. Not only can large numbers of cells be rapidly separated into cell cycle specific fractions, but the cells undergo very little manipulation, making this an extremely powerful tool for investigating gene expression at different stages of the cell cycle.

The following results are obtained using RNA from cell cycle sorted Human HeLa cells. The Mouse and Human Surfeit clusters have been shown to have the same organization of conserved genes. The Human Surf genes

homologous to Surf-1, Surf-2 and Surf-4 have been shown to be expressed by probing Northern blots of Human RNA with Mouse Surfeit probes (Williams et al 1988). A cosmid containing the human Surf-1, Surf-3 and part of the human Surf-2 genes has been isolated (J. Yon, unpublished results), showing that these genes have a similar organization in human as in mouse DNA.

6.2 Results

Two Northern blots (blot 1 and blot 2) of Human cytoplasmic RNA from HeLa cells were kindly provided by Dr. Morris. These HeLa cells had been separated into fractions enriched for different stages of the cell cycle by centrifugal elutriation. Each lane contains 10 µg of cytoplasmic RNA. Figure 6.1 shows these blots probed sequentially with chicken histone H2b (Krieg & Melton 1984), Surf-3 cDNA C4G, Surf-2 cDNA H1 and Surf-4 cDNA 15, cross hybridization to the 18S rRNA is also shown. The Surf-3 mRNA was easy to detect with a random primed DNA probe hybridized and washed at standard stringency (see materials and methods). However, Surf-2 and Surf-4 could not be detected under these conditions due to the low levels of the mRNAs in the cell, and the reduced homology due to using mouse probes on blots of human RNA. In order to increase the sensitivity, single strand RNA probes generated with SP6 polymerase were made and hybridized and washed at a variety of stringencies. Under conditions where Surf-2 and Surf-4 were detectable there was cross hybridization to the C+G rich rRNA and only the larger, 2.8 kb, Surf-4 mRNA was detected. Even when a single stranded RNA probe was used, Surf-1 could not be detected.

The level of histone H2b mRNA clearly changes throughout the cell cycle, being expressed at a much higher level during early S-phase than in G1 and G2, as would be expected (Stein et al 1984). The level of 18S rRNA is an internal control showing that the same amount of RNA is present in each lane and it does not show cell cycle variation in amount. To determine

whether the levels of Surfeit gene mRNAs change during the cell cycle, the autoradiographs were scanned with a laser densitometer and the area of each peak plotted as shown in figures 6.2, 6.3 and 6.4.

Figure 6.2 a) and b) show data provided by Dr. Morris. Figure 6.2 a) shows the percentage of cells in each of the ten fractions collected from the centrifugal elutriation, which are in the three cell cycle stages, G1, S, and G2 as determined by DNA content. The lane numbered 1 contains figures for the total population before centrifugal elutriation. Figure 6.2 b) shows the protein content and tritiated thymidine incorporation of the cells in each of the ten fractions from the centrifugal elutriation. Again, lane 1 is the values for the population of cells before centrifugal elutriation. It is clear that protein content increases during passage through the cell cycle and thymidine incorporation peaks in the eighth fraction which is also the fraction with the highest percentage of S phase cells. Figure 6.2 c) shows the percentage of cells of each fraction in S phase, and the results of the densitometric scan of one of the blots (blot 1) probed with chicken histone H2b. The level of histone mRNA clearly peaks in early S phase confirming that cell cycle sorting of the cells has taken place, and the levels of mRNA do reflect the situation in the cells at different stages of the cell cycle.

Figure 6.3 shows the areas of the peaks observed for each of the fractions when a single northern blot (blot 1) was probed with Surf-2 cDNA H1, Surf-3 cDNA C4G and chicken histone H2b. In each graph the areas of the peaks due to cross hybridization to 18S rRNA are shown as an internal control for this northern blot. The levels of Surf-2 mRNA show an increase in fraction 9 but this probably does not represent significant cell cycle regulation of Surf-2. The levels of Surf-3 mRNA closely follow the levels of 18S rRNA and no changes of level correlate with progression through the cell cycle. The level of histone H2b mRNA on this northern blot is also

shown, the levels are noticeably different to the levels of 18S rRNA, showing a peak in early S phase.

Figure 6.4 shows the results for Surf-4 and histone H2b which were obtained using a second northern blot (blot 2). Again, the 18S rRNA is used as an internal control. Only the larger 2.8 kb Surf-4 mRNA was detectable, this mRNA can be seen to follow the level of 18S rRNA closely, showing no cell cycle dependent changes in level. The histone mRNA, again, peaks in early S phase.

6.3 Discussion

Results obtained with human cytoplasmic RNA from cells sorted by centrifugal elutriation show that the levels of mRNA of Surf-3, Surf-2 and Surf-4 (2.8 kb mRNA) are not differentially regulated through the cell cycle. This does not rule out differential control of these genes at the translational level through the cell cycle, but it makes it unlikely that the layout of the genes is related to cell cycle regulation of transcription. Surf-3 has been identified as ribosomal protein L7a (Giallongo et al 1989) and as would be expected, the level of Surf-3 mRNA exactly follows the level of 18S rRNA in each RNA sample.

Figure 6.1 Northern blots of cell cycle dependent expression of the Surfeit genes.

Two Northern blots (blot 1 and blot 2) of cell cycle sorted HeLa RNA were kindly provided by Dr. Morris at Cold Spring Harbor. Each lane contains 10 μg of cytoplasmic RNA from one of 10 fractions of elutriated cells. The first lane on the left contains RNA from the unfractionated population, and the other lanes progress from G1 through to G2.

The Chicken histone H2B probe is a 650 bp PstI to EcoRI fragment from the plasmid pSP62-H2BΔ4 (obtained from David Bentley). This plasmid is derived from the plasmid pSP62-H2B, described in Krieg and Melton (1984), by a DraI to SmaI deletion (David Bentley personal communication).

The Surf-3 probe used was the EcoRI insert of cDNA C4G (see section 2.1) labelled by random oligonucleotide priming.

The Surf-2 probe was a continuously labelled SP6 RNA transcript of the cDNA H1. The Surf-4 probe was a continuously labelled SP6 RNA transcript of the cDNA 15, 928 bp EcoR1 fragment (see appendix 2a). For Surf-2 and Surf-4 hybridization was carried out at 64°C with 40% formamide and 400 mM NaCl. Washing was at 64°C and 40 mM $NaPO_4$.

The 18S rRNA signal was obtained from cross hybridization to the Surf-4 probe.

The histone H2B, Surf-3 and Surf-2 are all using northern blot 1 whereas the Surf-4 and 18S rRNA signal is on Northern blot 2.

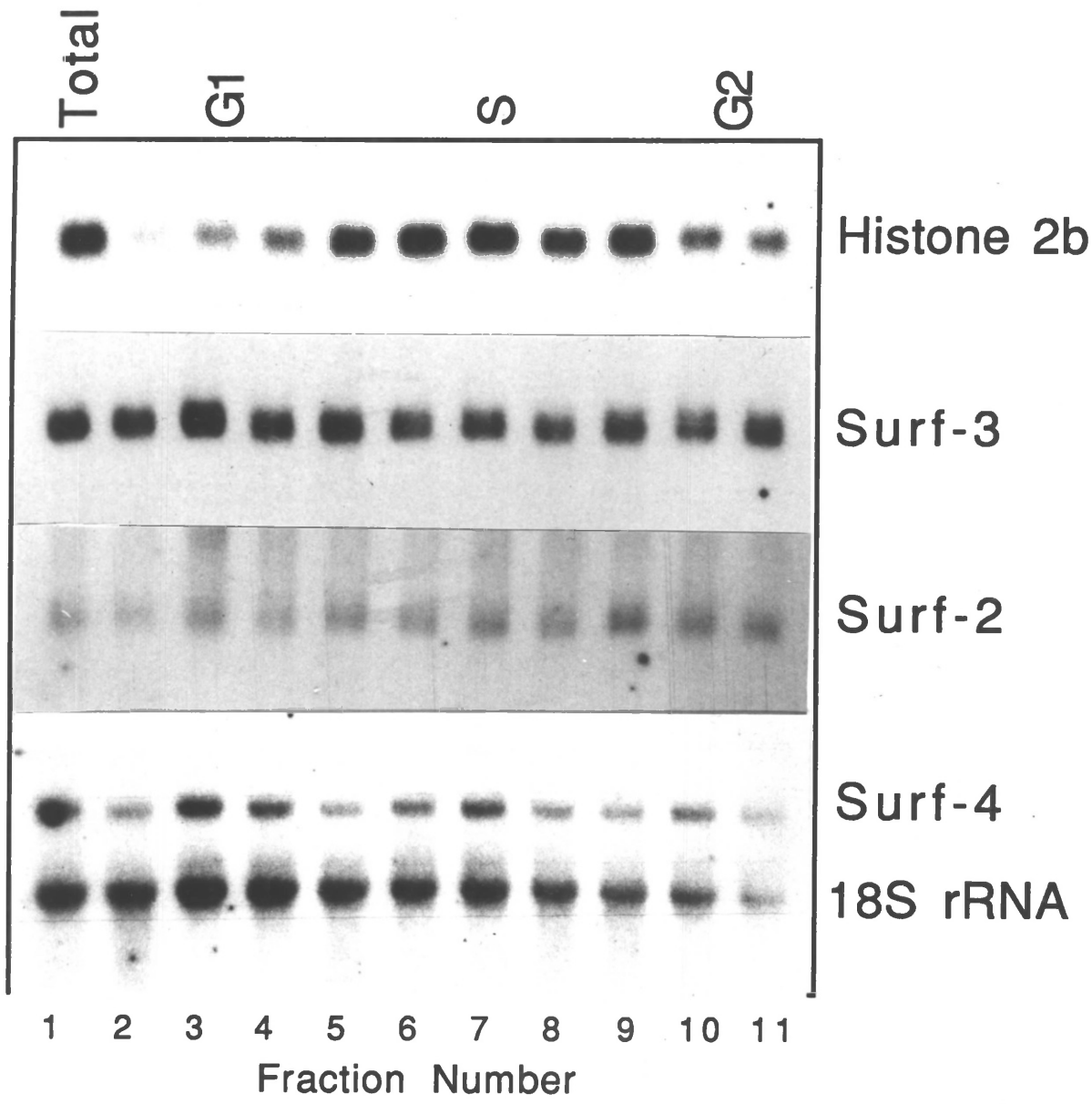

Total
G1
S
G2
Histone 2b
Surf-3
Surf-2
Surf-4
18S rRNA
1 2 3 4 5 6 7 8 9 10 11
Fraction Number

Figure 6.2 Graphs showing the cell cycle sorting of the HeLa cells.

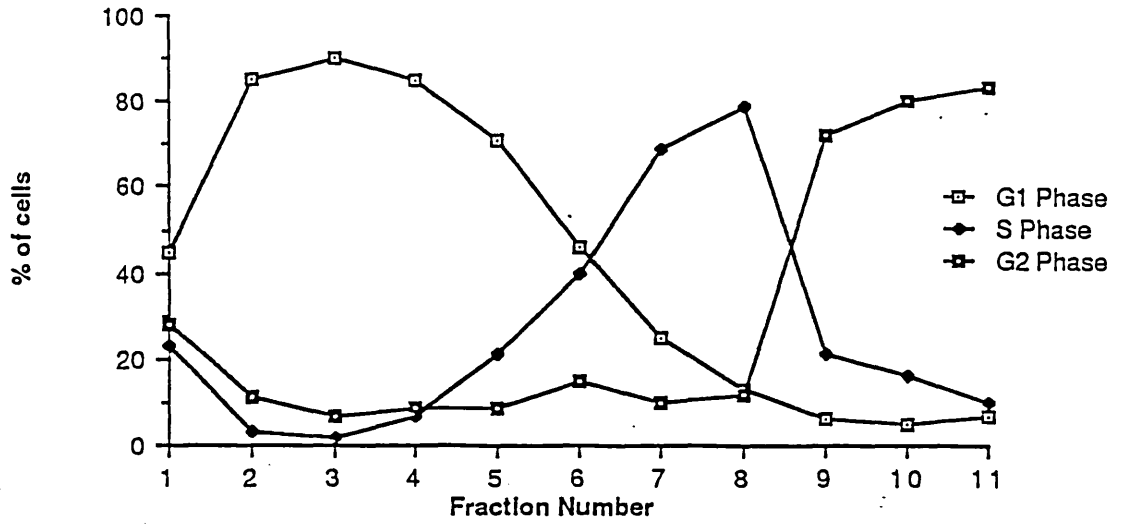
For all three graphs, lane 1 contains values for the total population of cells before centrifugal elutriation. Lanes 2 to 11 contain values for ten consecutive fractions collected from centrifugal elutriation.

a) Percentage of cells in each phase of the cell cycle. This was determined by FACS analysis (Data from Dr. Morris).

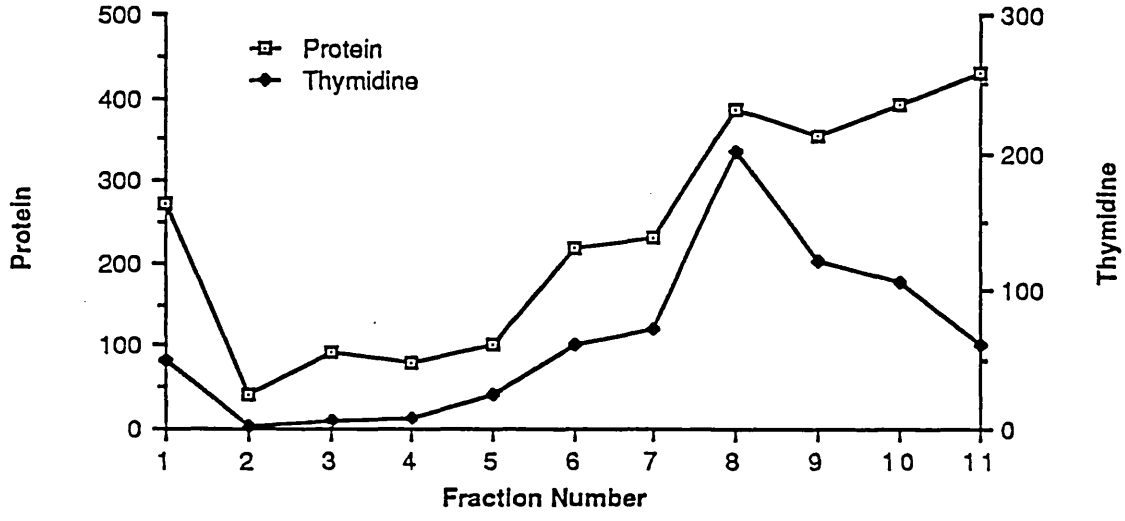
b) Protein content and tritiated thymidine incorporation. The left hand scale shows μg protein per 10^6 cells. The right hand scale shows incorporated tritiated thymidine cpm X 1000 per 10^6 cells (Data from Dr. Morris).

c) histone expression and percentage in S phase. The left hand scale shows the level of histone mRNA in arbitrary units. The level of mRNA is measured as the area under the peak of hybridization using a densitometer to scan the autoradiograph. The right hand scale shows the percentage of cells in S phase for each fraction of cells.

a) Cell Cycle Phase



b) Protein content & Thymidine Incorporation



c) Histone Expression & Percentage in S Phase

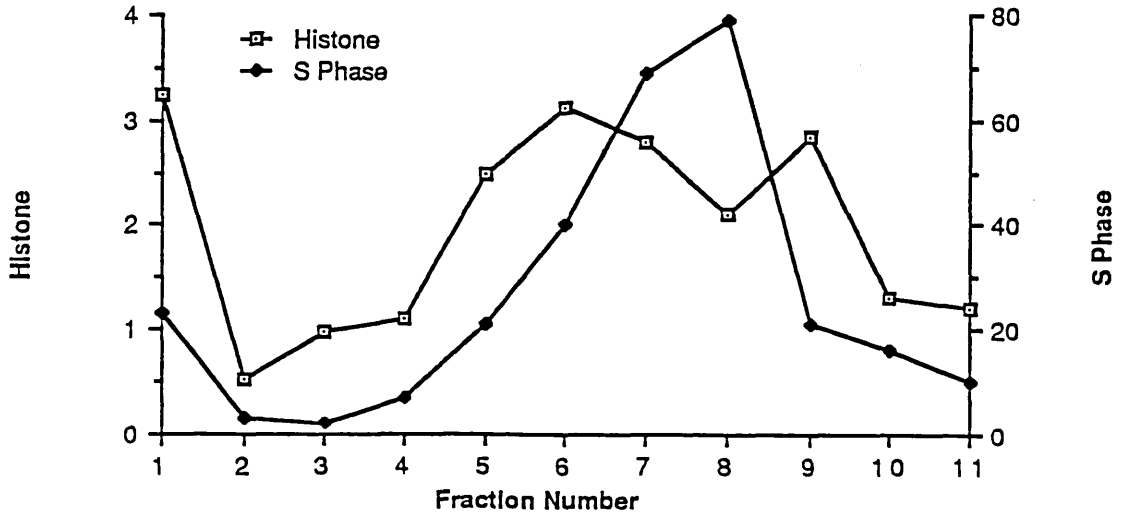
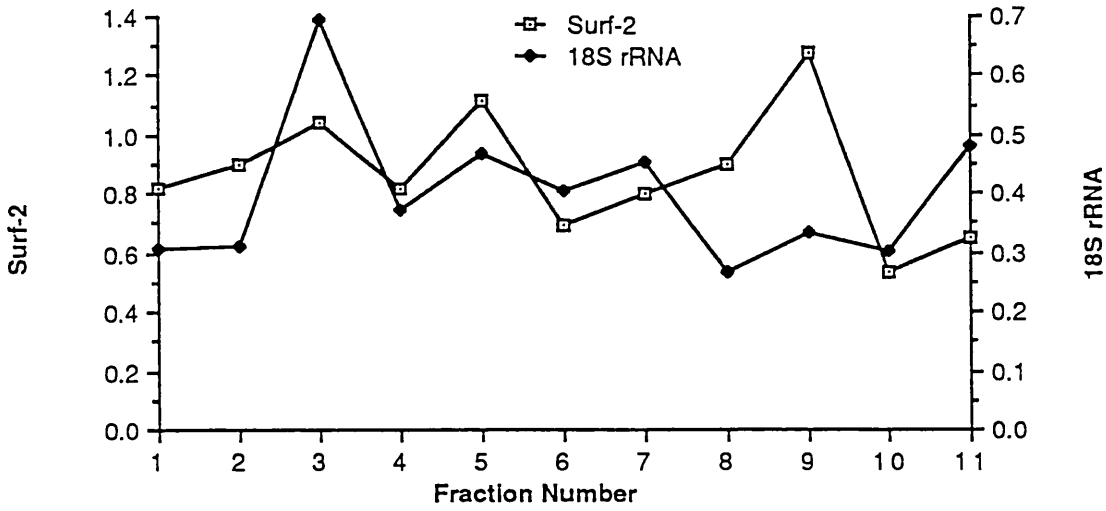


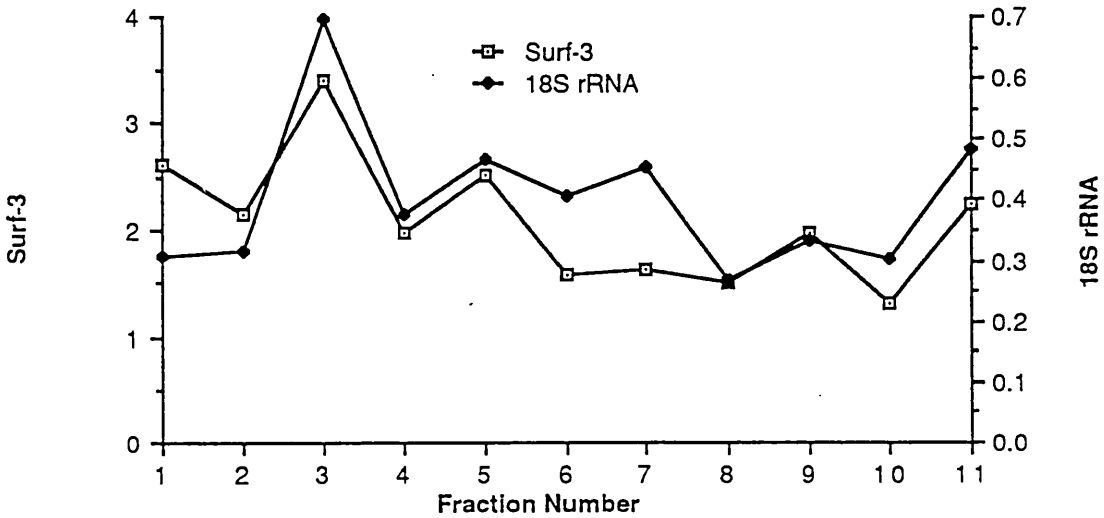
Figure 6.3 Expression of Surf-2 and Surf-3 through the cell cycle.

In each graph lane 1 contains values for the total population before separation by centrifugal elutriation. Lanes 2 to 11 contain values for consecutive fractions from a centrifugal elutriation. Lane 2 is the peak of G1, lane 8 is the peak of S phase and lane 11 is the peak of G2, as shown in figure 6.2a. Northern blot 1 was hybridized successively with Surf-2 cDNA H1, Surf-3 cDNA C4G, and chicken histone H2b (see figure 6.1 for probes and conditions). The Surf-2 cDNA H1 probe gave cross hybridization to the 18S rRNA and this has been used as an internal control for loading of the lanes. The autoradiograms of each hybridization were scanned with a densitometer and the areas of the peaks are plotted. The absolute area depends on the length of time of exposure of the autoradiogram and does not represent the level of expression of the mRNA.

Surf-2 mRNA & 18S rRNA



Surf-3 mRNA & 18S rRNA



Histone mRNA & 18S rRNA

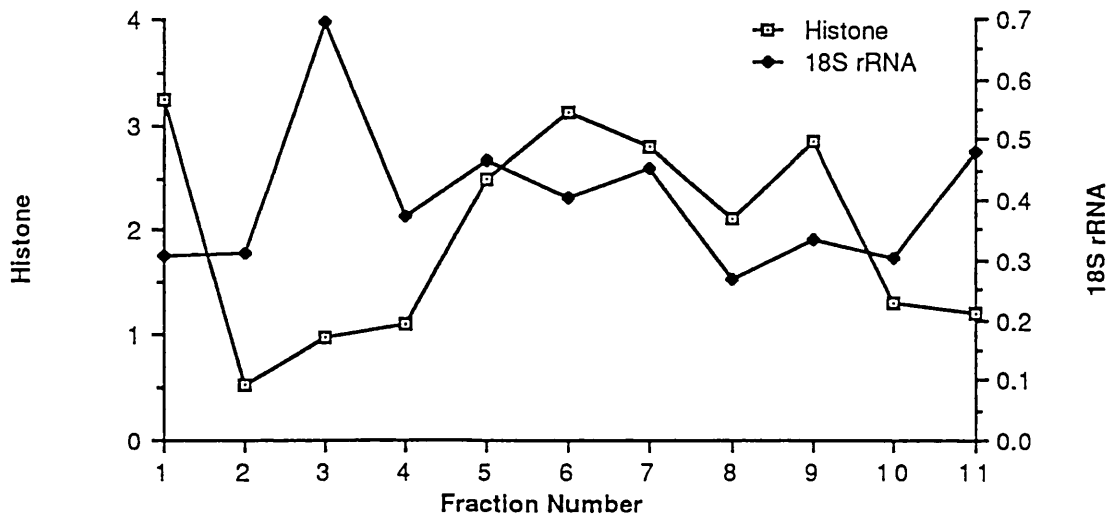
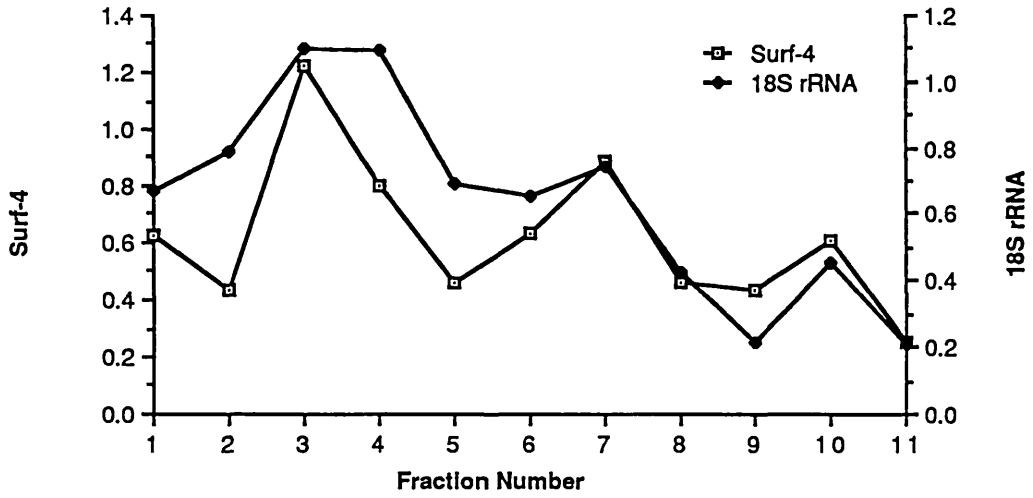


Figure 6.4 Expression of Surf-4 through the cell cycle.

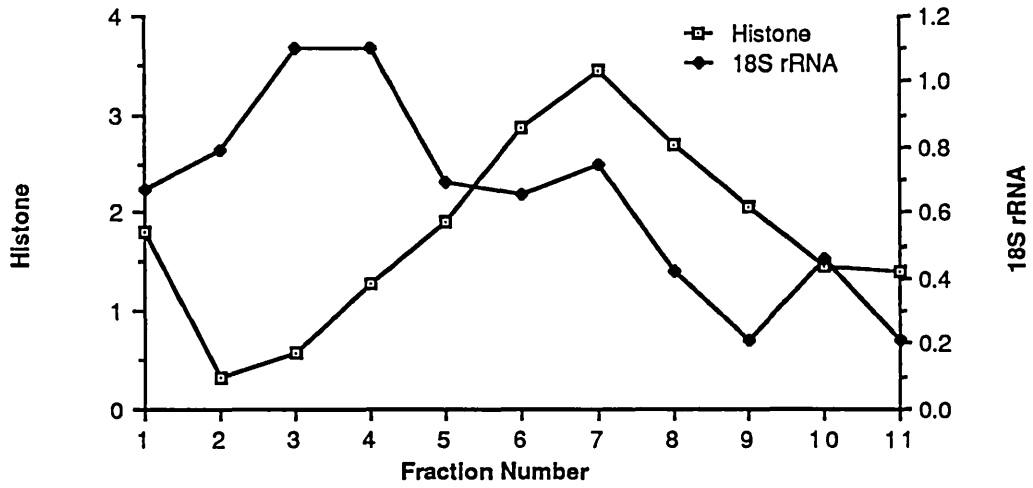
In each graph lane 1 contains values for the total population before centrifugal elutriation. Lanes 2 to 11 contain values for consecutive fractions from a centrifugal elutriation. Lane 2 is the peak of G1, lane 8 is the peak of S phase and lane 11 is the peak of G2, as shown in figure 6.2a. Northern blot 2 was hybridized successively with Surf-4 cDNA 15 (928 bp EcoRI fragment as a single stranded RNA probe, see figure 6.1 for conditions) and chicken histone H2b. The hybridization with Surf-4 gave cross hybridization to the 18S rRNA and this has been used for an internal control of the loading of the lanes. The autoradiograms of each hybridization were scanned with a densitometer and the areas of the peaks are plotted. The absolute area depends on the length of time of exposure of the autoradiogram and does not represent the level of expression of the mRNA.

The bottom graph shows the area under each peak after hybridization with the chicken histone H2b probe and the percentage of cells in S phase. The percentage of cells in S phase was determined FACS analysis by Dr. Morris.

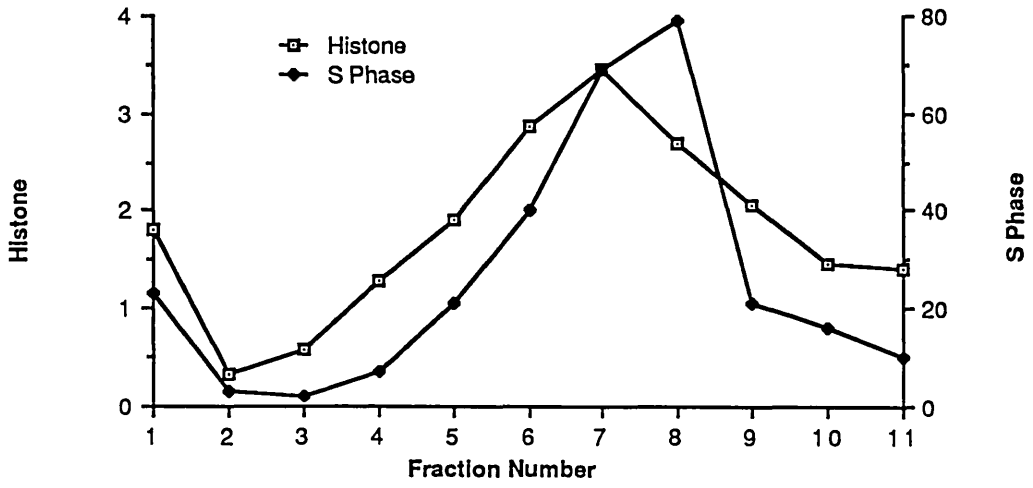
Surf-4 & 18S rRNA



Histone Expression & 18S rRNA



Histone Expression & Percentage in S Phase



Chapter 7: Mapping

7.1 Introduction

The Surfeit locus has an organization which is unique in the mammalian genome (see section 1.2). The unusual organization could be due to the region of the mammalian chromosome where it is located. For example it could be very near to a telomere where the genes might be more clustered. An example of very clustered CpG-rich islands which might be similar to the Surfeit cluster occurs near the telomere of human chromosome 4, near the Huntington's gene (see section 1.2) (Pohl et al 1988). To determine whether the Surfeit cluster is located in a particular region of the chromosome it has been mapped in both mouse and man.

The Surfeit gene cluster has been mapped to the tip of the long arm of human chromosome 9, bands 9q33-34, by hybridization to somatic cell hybrid DNA and by *in situ* hybridization (Yon et al 1989). This is consistent with its having a telomeric location.

The methods available to map genes in the mouse are *in situ* hybridization to metaphase spreads, hybridization to a panel of somatic cell hybrids, recombination of polymorphic markers in a panel of recombinant inbred strains, and recombination of polymorphic markers in a *Mus musculus/Mus spretus* interspecific cross (Avner et al 1988).

Hybridization to metaphase spreads is much harder with mouse chromosomes than with human chromosomes, because all the mouse chromosomes are acrocentric, they form a continuous gradation in size and they do not band well. If one already knows which chromosome the gene of interest is located on, then a Robertsonian translocation can be used to mark that chromosome.

Somatic cell hybrids between mouse and hamster cells rarely form cell lines containing a single mouse chromosome and very few cell lines are

available which contain fragments of mouse chromosomes so one can usually only map a gene to a whole mouse chromosome by this method.

Recombinant inbred strains have been formed by crossing homozygous inbred strains of mouse such as DBA/2 and C57Bl/6, the offspring of such a cross are then inbred for at least 20 generations, until the recombinant inbred lines are themselves homozygous (Taylor 1978). The resulting lines are each a patchwork of the parental genomes showing linkage between close genes. Small genetic distances can be measured but larger genetic distances are difficult to detect using this system. In order to map a gene using recombinant inbred strains it is necessary to have a probe which detects a polymorphism between the two parental strains. Even though the commonly used laboratory strains of mice are derived from different subspecies of *M. musculus* it is often difficult to find such a polymorphism.

The problem of lack of variability is overcome by using the interspecific cross between *M. musculus* and *M. spretus* (Robert et al 1985). *M. spretus* is used because it is genetically distant from *M. musculus* (Bonhomme et al 1984), and it gives fertile females in the F1 generation although the F1 males are infertile (Avner et al 1988). The F1 females are backcrossed to the parental *M. musculus* strain and the F2 offspring are analysed for inheritance of polymorphic markers. Polymorphisms between *M. musculus* and *M. spretus* are normally common and genes can be mapped over a range of genetic distances up to about 30 centiMorgans (cM). It has been shown that genetic distances between markers on the X chromosome in this interspecific cross are indistinguishable from distances found with crosses within *M. musculus* (Avner et al 1988). However there may be regions of the genome which are different between *M. musculus* and *M. spretus* and recombination in these regions will not represent genetic distances in *M. musculus*. In particular, inversions such as the t-complex on chromosome 17 (Herrmann et al 1986) will prevent recombination in these

regions in the interspecific cross.

7.2 Results: Mapping the Surfeit Cluster and c-*abl*

To map the Surfeit locus using the interspecific cross between *M. musculus* and *M. spretus*, a search for polymorphisms between the two parental species was carried out with filters kindly provided by Dr. Mya Bucan. A single polymorphism was detected in HindIII digested DNA when probed with the XhoI to HindIII fragment at the far left of the 19 kb EcoRI fragment (see appendix 3).

The c-*abl* gene maps close to the Surfeit cluster at the tip of the long arm of human chromosome 9, and has also been mapped by *in situ* hybridization to band B of mouse chromosome 2 (Threadgill & Womack 1988). In order to map the mouse c-*abl* gene more accurately, a v-*abl* probe, previously characterized by Dale and Ozanne (1981) was used to search for polymorphisms between *M. musculus* and *M. spretus*. The 1.9 kb SacI - HindIII fragment used hybridizes to unique mouse fragments corresponding to the c-*abl* gene. A polymorphism in BglII fragment size was detected.

Two other markers used in this study were *β2-microglobulin* and *Hox*-5.1. *β2-microglobulin* is located about 40 cM from the centromere on mouse chromosome 2 (Lyon 1988b). *Hox*-5.1 is a murine homeo box-containing gene located by *in situ* hybridization to band D of mouse chromosome 2 (Featherstone et al 1988).

Two sets of F2 animals from interspecific crosses using different strains of mice were used. The first set consists of 34 animals which have been described elsewhere (Bucan et al 1986). The distribution of DNA polymorphisms for *β2-microglobulin*, *Hox*-5.1, c-*abl* and the Surfeit cluster are shown in Table 7.1. The second set consists of 54 animals which have not been previously published. The distribution of DNA polymorphisms in these animals is shown in Table 7.2. The genetic distances derived from both

crosses are shown in Table 7.3.

The two sets of backcross mice are derived from different parental strains so it is possible that there are actual differences between the recombination frequencies in the two crosses. However, the error due to sampling is reduced if the animals of both crosses are pooled and the genetic distances for all the animals combined are also given in Table 7.3. The genetic distances calculated from all the animals combined are; c-*abl* to *Surf* = 6±3 cM, *Hox*-5.1 to c-*abl* = 20±5 cM and *β2-microglobulin* to *Hox*-5.1 = 19±5 cM. The set of 37 backcross mice have been scored for the anchor locus *Agouti* and this is located a further 20±5 cM from *β2-microglobulin*. As *β2-microglobulin* is proximal to *Agouti*, this indicates that the Surfeit cluster, c-*abl* and *Hox*-5.1 are proximal to *β2-microglobulin*. Figure 7.1 Shows a map of mouse chromosome 2 with the positions of the markers mapped in this study and some other markers previously mapped to this region.

A combination of previous results indicate that *Agouti* and *β2-microglobulin* are separated by 18 cM (Lyon 1988b) and this corresponds well with the value of 20±5 cM in the set of 34 interspecific backcross animals indicating that recombination in this region is not distorted. About 40 markers are located on a recently published map of chromosome 2 between *Agouti* and the centromere (Lyon 1988b). *Sd* is the most proximal previously mapped marker and the distance between *Agouti* and *Sd* is about 55 cM. This study indicates that the Surfeit cluster is 65 cM from *Agouti* but the error on this value is quite large. Thus the Surfeit locus may be the most centromeric known marker on chromosome 2. However the error in the genetic distances is such that *Sd* could be on either side of the Surfeit cluster and c-*abl*. The marker next to *Sd* is *stb* and this is located 11 cM distal to *Sd*, so *stb* is probably distal to both c-*abl* and the Surfeit cluster.

The use of recombinant inbred strains to map genes, allows one to determine smaller genetic distances more accurately than the interspecific

cross. This is because the formation of recombinant inbred strains corresponds to approximately four meiotic recombination events rather than the single event in the interspecific cross (Taylor 1978). Blots of genomic DNA from C57 and DBA mice, the most commonly used parental strains, restricted with a wide variety of different restriction enzymes were kindly provided by Dr. Mya Bucan and Dr. Lisa Stubbs. The probes used to detect polymorphisms were: Surf-3 intron probe (appendix 1a), Surf-1 cDNA T1B, Surf-2 cDNA H1, Surf-4 cDNA 15, and the XhoI to HindIII fragment from the left of the 19 kb EcoRI fragment (appendix 3). The enzymes used were TaqI, BglII, BamHI, EcoRI, HindIII, PvuII, MspI, XbaI, and KpnI. None of the probes used detected a polymorphism with any of the enzymes indicating that the whole of the Surfeit cluster is highly conserved as would be expected of a region covered with housekeeping genes. In addition the v-*abl* probe was also used to attempt to detect polymorphisms between C57 and DBA, again, no polymorphisms were detected with AccI, HincII, HpaI, KpnI, NdeI, PstI, SstI, XmnI and XbaI, so this method of mapping could not be used. The v-*abl* probe only contains the 3' exons of c-*abl* which are spread over about 20 kb, it does not span the very large first intron of c-*abl* (Bernards et al 1988). Probes derived form the large first intron may detect polymorphisms between inbred strains.

Table 7.1 Distribution of DNA polymorphisms in an interspecific backcross of 34 animals.

Female F1 progeny from a cross between B6CARB.I. and SPE were backcrossed with B6CARB.I. to give F2 animals which were analysed (Bucan et al 1986). B6CARB.I. is a *Mus. m. domesticus* recombinant inbred strain between CBA and C57BL/6, SPE is is a *M. spretus* strain inbred for over 25 generations (Bucan et al 1986). Filters carrying restriction enzyme digested DNA from 37 F2 animals were a kindly provided by Dr. Lisa Stubbs. The data for *β2-microglobulin*, and *Hox*-5.1 were provided by Dr. Lisa Stubbs (unpublished results). Hybridization was carried out as described in materials and methods. + indicated the presence of, and - the lack of, the *M. spretus* allele. All the mice carry at least one of the *M. musculus* allele.

F2 animal	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
β2-mic.	-	-	+	?	-	+	-	-	-	?	-	-	+	?	+	+	?
Hox-5.1	-	-	+	-	-	+	-	+	-	+	-	-	+	+	-	-	-
c-abl	-	-	+	?	+	+	-	-	-	-	-	-	?	+	-	-	+
Surfeit	-	-	?	-	+	+	-	-	-	-	-	-	+	+	-	-	+

F2 animal	27	28	29	30	31	32	34	35	36	37	38	39	40	41	42	43	44
β2-mic.	+	-	-	+	-	+	-	+	-	+	-	+	+	+	+	-	-
Hox-5.1	+	-	-	+	-	+	-	-	+	+	-	+	+	?	-	-	?
c-abl	+	-	-	+	+	?	+	-	+	?	-	+	?	+	-	-	?
Surfeit	+	-	-	+	+	-	+	-	+	?	-	+	+	+	-	-	?

Table 7.2 Distribution of DNA polymorphisms in an interspecific backcross of 54 animals

The small eye (Sey) mutation has been crossed onto C57Bl/6 for 7 generations, and this strain used for a backcross with *M. spretus* (Dr. Lisa Stubbs personal communication). Filters with restricted DNA from individual animals from this backcross were the kindly provided by Dr. Lisa Stubbs. The data for *β2-microglobulin*, and *Hox*-5.1 are from Lisa Stubbs (unpublished results). Hybridization was carried out as described in materials and methods. + indicated the presence of, and - the lack of, the *M. spretus* allele. All the mice carry at least one copy of the *M. musculus* allele. The probes used for *c-abl* and the Surfeit cluster are described in section 7.2.

F2 animal	A1	C1	E1	G1	I1	A2	C2	E2	G2	I2	A3	C3	E3	G3	I3	A4	C4	E4	G4	I4
β2-mic	+	+	+	-	?	+	+	+	+	+	+	-	+	+	-	+	+	-	?	-
Hox-5.1	+	+	+	-	?	+	+	+	+	+	+	-	+	-	+	+	+	+	+	-
c-abl	+	-	?	?	-	?	+	+	?	+	+	?	?	-	?	+	+	+	+	-
Surfeit	+	-	+	-	-	+	+	+	+	-	+	+	-	-	+	+	+	+	+	-

F2 animal	A5	C5	E5	G5	I5	A6	C6	E6	G6	I6	A7	C7	E7	G7	I7	A8	C8	E8	G8	I8
β2-mic.	+	?	-	+	+	+	+	-	-	+	+	+	-	+	+	-	-	+	+	+
Hox-5.1	-	+	-	+	+	-	-	-	-	+	+	+	-	+	+	-	-	+	+	+
c-abl	+	?	-	+	-	-	-	-	-	+	+	+	+	-	+	-	-	+	-	+
Surfeit	+	+	-	-	+	-	-	-	-	+	+	+	+	-	+	-	-	+	-	+

F2 animal	A9	C9	E9	G9	I9	A10	C10	E10	G10	I10	A11	C11	E11	A12
β2-mic.	+	+	+	+	+	-	-	-	-	-	?	?	?	?
Hox-5.1	-	-	+	+	+	-	-	-	-	-	-	?	+	-
c-abl	-	-	+	+	+	-	+	-	-	?	-	+	+	+
Surfeit	-	-	+	+	+	-	+	-	-	-	-	+	-	+

Table 7.3 Genetic distances between markers

1st marker	2nd marker	M/N	cM	standard error
34 backcross animals				
β2-mic	Hox-5.1	6/28	21	±8
Hox-5.1	c-abl	6/27	22	±8
c-abl	Surfeit	0/27	0	<4
54 backcross animals				
β2-mic	Hox-5.1	8/47	17	±5
Hox-5.1	c-abl	8/43	19	±6
c-abl	Surfeit	4/45	9	±4
Combined results				
β2-mic	Hox-5.1	14/75	19	±5
Hox-5.1	c-abl	14/70	20	±5
c-abl	Surfeit	4/72	6	±3

The figures for the "34 backcross animals" are from table 7.1, and those for the "54 backcross animals" are from table 7.2. For the "combined results" the numbers of animals from the two backcrosses are pooled.

M = number of recombinants

N = number of animals

p = M/N

When p x N > 5, the distribution of p is an approximation to the normal distribution, and the standard error of p, $\sigma p = \sqrt{p(1-p)/N}$

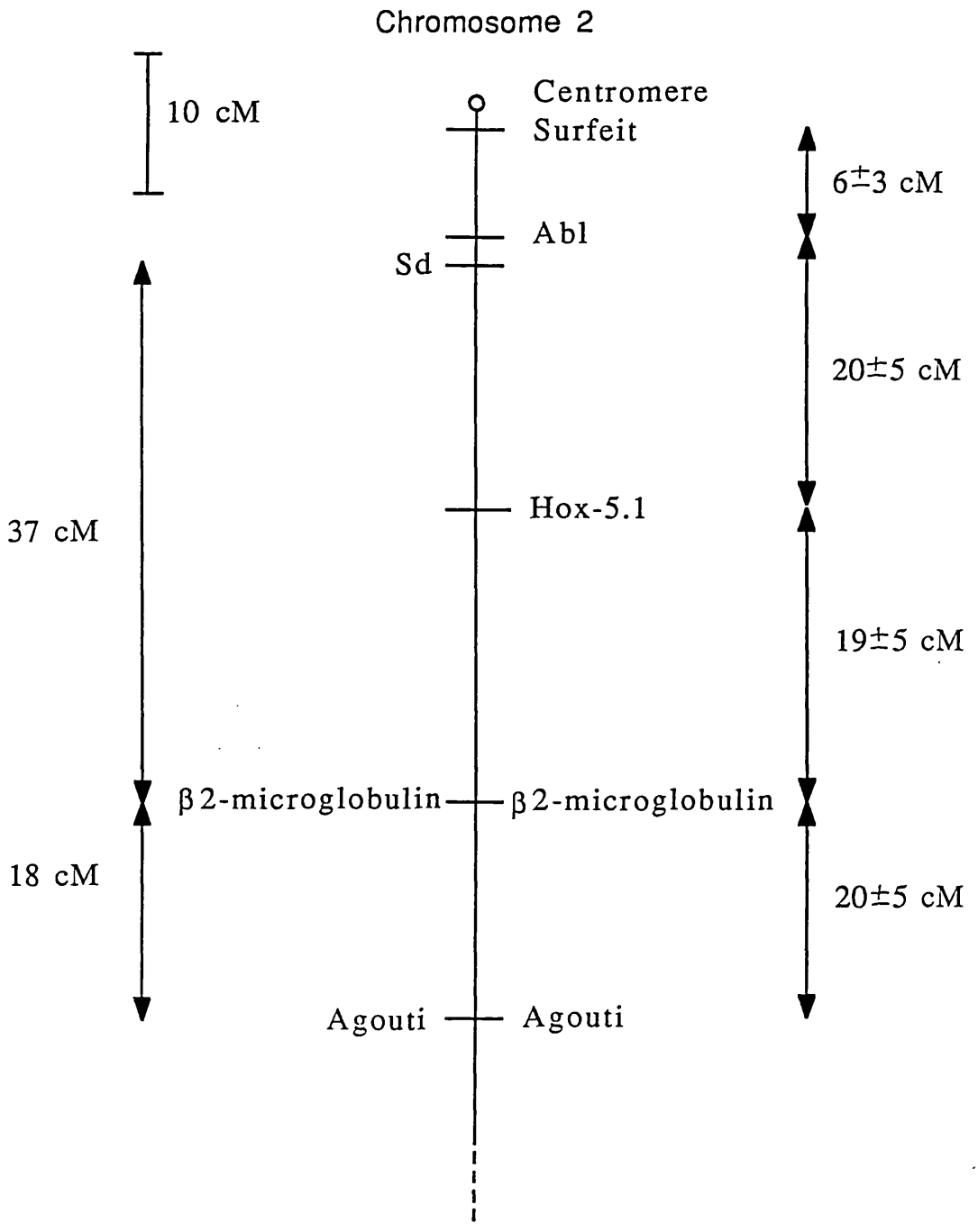
The approximation to the normal distribution cannot strictly be used to estimate the standard error of p for c-*abl* to the Surfeit cluster. The above formula has however been used to estimate this standard error.

Genetic distance is measured in centiMorgans (cM) where 1 cM = 100 x p and the standard error of the number of cM = 100 x σp

There is a 5% chance of the real genetic distance being outside the calculated value ± 2 x the standard error. (Glantz 1987).

Figure 7.1 Map of mouse chromosome 2

The vertical line represents the proximal half of mouse chromosome two with the centromere at the top. The genetic markers *Agouti*, *β2-microglobulin*, *Hox-5-1* and the Surfeit cluster, which have been mapped in this study are indicated with the genetic distances shown on the right. The *Sd* gene, which has not been cloned, but is the most proximal previously identified marker on chromosome 2, is also indicated. The marker *Sd* could be located on either side of the Surfeit cluster and c-*abl* because of errors in the genetic distances. The genetic distances from combined previous work (Lyon 1988b) are shown on the left.


Chromosome 2
10 cM
Centromere
Surfeit
Abl
Sd
6±3 cM
20±5 cM
Hox-5.1
37 cM
19±5 cM
β2-microglobulin
β2-microglobulin
18 cM
20±5 cM
Agouti
Agouti

7.3 Results: The Surfeit Cluster and c-*abl* Cannot be Linked With Pulsed Field Gel Electrophoresis

With the set of 37 backcross animals (see section 7.2) no recombinants were observed between the c-*abl* gene and the Surfeit cluster, suggesting that c-*abl* and the Surfeit cluster are very closely linked, though with the set of 54 animals several recombinants were observed. In the mouse, 1 cM corresponds very roughly to 2 mega base pairs of DNA (Mb) (Barlow & Lehrach 1987). On pulsed field gel electrophoresis one can separate DNA fragments up to 4 Mb in size so it was possible that the c-*abl* gene would lie in the same DNA fragment as the Surfeit cluster, if large enough fragments were analysed. A pulsed field gel was run with mouse DBA DNA digested with MluI, NotI, SfiI, PvuI, NruI and RsrII under conditions which resolve fragments from 300 kb to 3000 kb in size. NotI cuts mammalian genomic DNA into fragments on average 1,000 kb in size, MluI, NruI, and PvuI give fragments on average 300 kb in size and SfiI and RsrII give fragments on average 200 kb in size (Appendix to Biolabs catalogue 1988). Moreover these enzymes (except for NotI) are not specific for CpG-rich islands, which is important as I had already demonstrated the presence of the CpG-rich island specific enzyme cutting sites 500 kb and 100 kb on either side of the Surfeit cluster (see section 4.3).

The gel was blotted and the filter probed sequentially with a fragment from the far left of the Surfeit cluster (see figure legend), a fragment from the far right of the Surfeit cluster (see figure legend) and v-*abl* as shown in figure 7.2. A clear NotI fragment of approximately 700 kb is seen for both the surfeit cluster probes indicating that the cluster does not itself contain any NotI sites. The v-*abl* probe detects a NotI fragment that is larger than the Surfeit cluster NotI fragment. Various large PvuI, NruI and RsrII fragments are detected with the Surfeit probes and the v-*abl* probe but in no instances are the same fragments detected by the v-*abl* probe as by the Surfeit probes

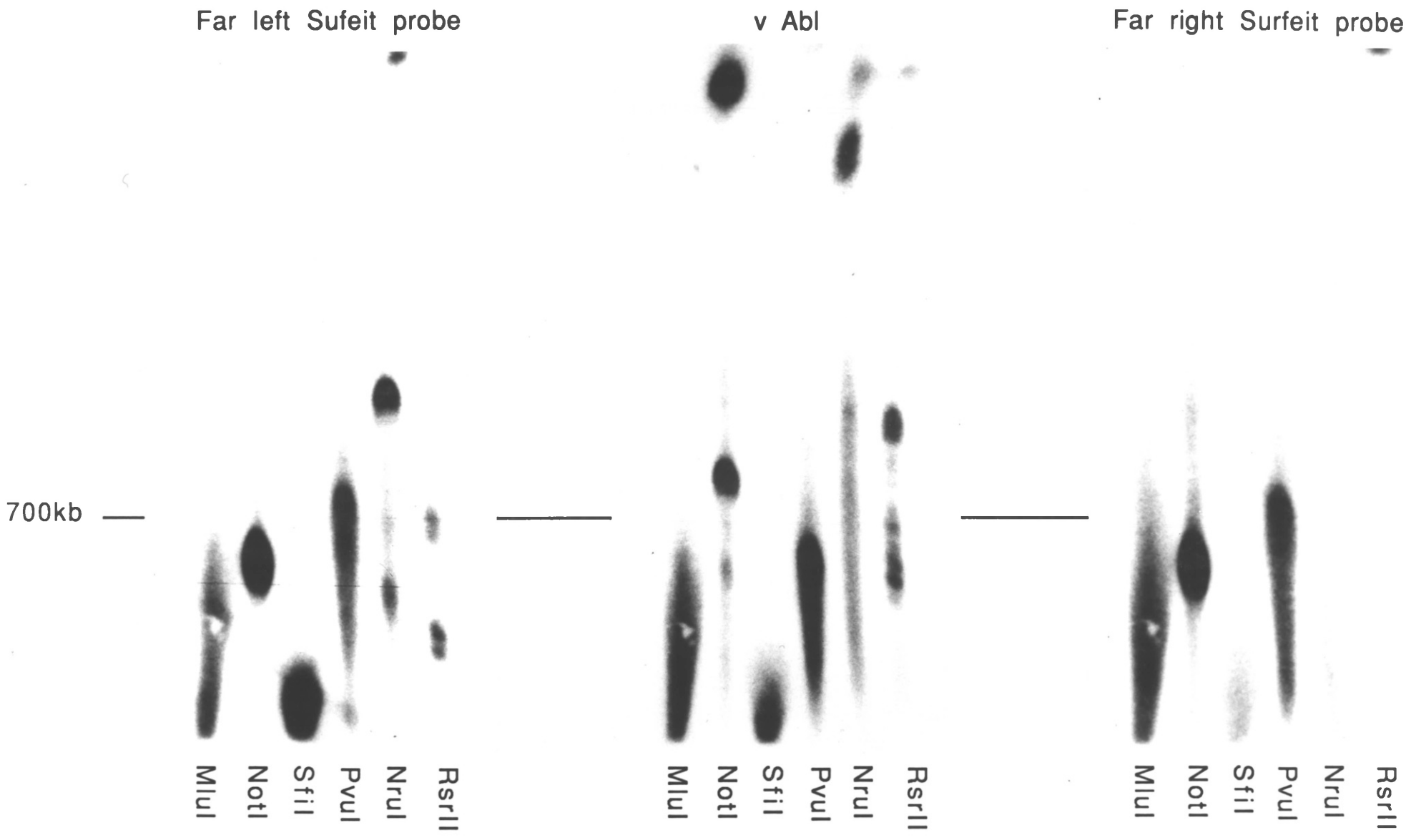
indicating that none of the rare cutting enzymes give fragments that span both the Surfeit cluster and c-*abl*. However, one cannot specify a minimum distance between the Surfeit cluster and c-*abl* because the fragments detected could extend to either side of the probes used.

The mouse c-*abl* gene has a first intron spanning at least 70 kb (Bernards et al 1988). The human first intron is at least 200 kb in size (Bernards et al 1987) indicating that the mouse gene might also have a first intron spanning hundreds of kilobases. The v-*abl* probe used here contains only the 3' end of c-*abl* and a probe at the 5' end of c-*abl* might detect fragments in common with the Surfeit probes.

Figure 7.2 Pulsed field gel hybridized with Surfeit probes and v-*abl*

Low melting agarose blocks containing $1x10^6$ Balb/C A31 mouse cells were prepared and digested with the rare cutting enzymes as described in materials and methods. The pulsed field gel was run at 170 V for 42 hrs with a switching time of 100 secs, these conditions resolve fragments which are 300 kb to 3000 kb in size. The yeast chromosomes used as markers for this gel were not visible but the NotI fragment detected with the Surfeit probes is approximately 700 kb in size as judged by other pulsed field blots where concatermerized lambda DNA was used as markers.

The DNA was transferred to Genescreen plus as described in the materials and methods. The filter was sequentially hybridized with the Surfeit and v-*abl* probes. The "Far left Surfeit probe" is the XhoI to HindIII fragment from the far left of the 19 kb EcoRI fragment (see appendix 3). The "vAbl" probe is the 1.9 kb SacI - HindIII fragment containing the 3' half of v-*abl* sequences of Abelson murine leukemia virus fragment described by Dale & Ozanne (1981). The "Far right Surfeit probe" is the 800 bp ApaI fragment located upstream of Surf-4 (see appendix 2a).


Far left Sufeit probe
v Abl
Far right Surfeit probe
700kb
MluI
NotI
SfiI
PvuI
NruI
RsrII
MluI
NotI
SfiI
PvuI
NruI
RsrII
MluI
NotI
SfiI
PvuI
NruI
RsrII

7.4 Discussion

Using the interspecific cross between *M. musculus* and *M. spretus*, the Surfeit cluster has been mapped very close to the centromere of chromosome 2, approximately 45 cM from *β2-microglobulin*, approximately 26 cM from *Hox*-5.1 and 6±3 cM from c-*abl*. The Surfeit cluster has also been mapped to the tip of the long arm of human chromosome 9 (Yon et al 1989), so it is possible that the organization of the Surfeit cluster is characteristic of the telomeric and centromeric regions of chromosomes. The human c-*abl* gene is located with the Surfeit cluster on the tip of the long arm of chromosome 9 and there are several other genes which have been mapped to the same region of human chromosome 9 and mouse chromosome 2 (though they have not been positioned on mouse chromosome 2). Thus the Surfeit locus and c-*abl* form a large syntenic group which includes the genes adenylate kinase-1 (*Ak*-1), arginosuccinate synthetase (*Ass*), non-erythroid alpha-spectrin-1 (*Spna*-1), and folypolyglutamate synthetase (*Fpgs*) which are located together in both the mouse and human genome (Lalley et al 1988). *Hox*-5.1 has not been mapped in man. The human *β2-microglobulin* gene is located on human chromosome 15q11-q26.

It was not possible to map the Surfeit cluster more accurately using recombinant inbred strains because no polymorphisms were found between the parental strains, C57 and DBA for probes spanning the entire cloned Surfeit cluster and the v-*abl* probe. As the c-*abl* gene has an extremely long first intron which is not included in the v-*abl* probe (Bernards et al 1988), it is possible that probes from this intron might detect polymorphisms. Finding a polymorphism among the inbred strains would allow us to map the Surfeit cluster much more accurately.

The *Sd* mutation causes a condition called Danforth's short tail. Heterozygotes have short tails with a reduced number of caudal vertebrae and some kinking (Green 1981). *Sd* is the closest marker to the Surfeit cluster

and it would be useful to map the Surfeit cluster with respect to *Sd*. As *Sd* has not been cloned and the Surfeit cluster only detects polymorphisms between the two different mouse species *M. musculus* and *M. spretus*, it would be necessary to set up a cross between a *M. spretus* strain and a *M. musculus* strain carrying *Sd*.

As described in the introduction (section 1.2), it is a feature of higher eukaryotic genomes that genes which are functionally related but evolutionarily distinct are not clustered together, the exceptions being histone genes and rRNA genes. Surf-3 has been identified as an actively transcribed gene for ribosomal protein L7a (Giallongo et al 1989), and it is of interest to know whether this group of evolutionarily diverse but co-regulated proteins is clustered in the same way as the histone and rRNA genes. Although numerous mammalian ribosomal protein cDNAs have been cloned, only five functional genes (other than Surf-3) have been isolated due to the difficulty of distinguishing expressed genes from processed pseudogenes (see section 2.4). Three of these, L32, L30 and S16 have been mapped to mouse chromosomes 6, 15 and 7 respectively (Wiedemann et al 1987). Clearly none of these are located with Surf-3 on chromosome 2.

Chapter 8: Conclusion and Discussion

8.1 Summary of Results

This thesis presents investigations into the organization, identity and expression of the genes in the Mouse Surfeit gene cluster. The Mouse Surfeit gene cluster has previously been identified as a very tight cluster of at least 4 genes with an organization which is unique amoung known mammalian genes (section 1.2) (Williams et al 1988). The genes, called Surf-1 to -4 are organized as closely apposed alternating transcription units covering about 24 kb of genomic DNA. The 3' ends of two genes overlap by 133 bp and the 3' ends of two others are separated by 70 bp. Two genes are transcribed from a bidirectional promoter with their heterogeneous 5' ends separated by 15-73 bp. These features suggest that the organization is related to the control of expression of these genes.

Two genomic clones and cDNAs covering the Surf-4 gene have been cloned and analysed. This gene is arranged as 6 exons spanning 14 kb of DNA and encodes a putative protein of 269 amino acids. The two Surf-4 mRNA species differ in their use of poly(A) addition signal such that the 2.8 kb but not the 2.0 kb mRNA overlaps Surf-2.

The sequencing of Surf-3 cDNAs and genomic exons has been completed showing that Surf-3 encodes a highly basic (24% Arg+Lys) protein. This brings the number of characterised Surfeit genes to four. None of these genes are related by DNA or protein sequence identity. The Surf-1, -2 and -4 genes have not been identified, the DNA sequences have no homology to sequences in the EMBL or GENBANK databases and the protein sequences have no homology to proteins in the PIR database.

The upstream region of the Surf-3 and surf-4 genes have been sequenced and the start of transcription determined by primer extension and S1 mapping. All three characterised upstream regions in the Surfeit cluster

(Surf-1/-2, Surf-3 and Surf-4) are located in CpG-rich islands. They are C+G rich and do not contain TATA or CAAT boxes which is characteristic of ubiquitously expressed "housekeeping" genes. However, none of the three characterised upstream regions are related by DNA sequence identity and they do not contain the same transcription factor consensus binding sites, though an SP1 binding site occurs upstream of both Surf-3 and Surf-4.

Surf-3 is a highly expressed member of a multigene family and by cloning and sequencing a number of the other copies from mouse DNA, most if not all of these other genes have been shown to be pseudogenes formed by reverse transcription and integration.

Surf-3 has subsequently been shown to be the gene encoding ribosomal protein L7a (Giallongo et al 1989). However, the other Surfeit genes are probably not ribosomal protein genes as they are expressed at low levels and are single copy genes. The sequence of the 5' upstream region of Surf-3 has been compared with the five other known mammalian ribosomal protein genes. The two features in common between these upstream regions are the presence of a CpG-rich island and a pyrimidine tract spanning the CAP site. There is no evidence of a common sequence which could bind a transacting transcription factor which would coregulate ribosomal protein genes.

Two new genes, Surf-5 and Surf-6 have been identified in the Surfeit cluster bringing the total number of genes to six.

Four CpG-rich islands have been identified in the Surfeit cluster occuring at the 5' end of each gene. However, two other mammalian ribosomal protein genes are shown not to be located in clusters of CpG-rich islands, indicating that clusters similar to the Surfeit locus are not a feature of ribosomal protein genes.

Using the presence of CpG-rich islands as a method of detecting all "housekeeping" and most other genes, it appears that we have come to the end of the Surfeit Cluster, the neighbouring islands being 100 kb and 500 kb

on either side of the cluster. This indicates that the Surfeit cluster of six genes is a local phenomena being surrounded by "normally" spaced islands (section 1.2).

The surfeit cluster has been mapped using a *M. musculus/M. spretus* interspecific cross. It is located on mouse chromosome 2, approximately 6 cM proximal to c-*abl* and 45 cM proximal to *β2-microglobulin*, making it possibly the most proximal marker on mouse chromosome 2. The Surfeit cluster, c-*abl* and four other genes are all located together on mouse chromosome 2 and human chromosome 9. The Surfeit cluster is very close to the centromere of mouse chromosome 2 and near the telomere of human chromosome 9, so the organization of the Surfeit genes may be a feature of the ends of chromosomes. Surf-3 is not located on the same chromosome as any previously mapped ribosomal protein gene.

The steady state levels of Surf-2, -3 and -4 mRNAs have been found not to be differentially regulated during the cell cycle indicating that the organization of the genes is not related to this type of temporal control.

8.2 The Surfeit Cluster is a Novel Arrangement of Genes

As described in the introduction, most mammalian genes are widely separated on the genome (section 1.2). The exceptions to this are 1) clusters of related genes formed by relatively recent gene duplication, 2) the bidirectional promoters which have recently been observed to be associated with CpG-rich islands and 3) rare occurrences of close or overlapping unrelated genes.

The Surfeit genes bear no relationship to each other at the DNA sequence or protein sequence similarity level and so are not recently duplicated genes. At least one, and possibly two other, of the four CpG-rich islands in the Surfeit cluster act as bidirectional promoters (Williams & Fried 1986a), so the Surfeit cluster falls into the class of CpG-rich islands which act

as bidirectional promoters. It will be interesting to see whether these bidirectional promoters share features such as coregulation of the flanking genes as opposed to alternate expression, and whether they share the same transcription factors. However, none of the other described bidirectional promoters have been found in clusters of more than two genes. Indeed no cluster of genes containing so many unrelated genes in such a small amount of DNA, or with such little DNA separating adjacent genes, has been observed.

The Surfeit gene cluster could be a recent addition to the mammalian genome having originated from a more compact genome. Alternatively the Surfeit cluster could be a normal arrangement of mammalian genes, simply more crowded than previously observed, but the crowding having no particular significance. Or, the Surfeit cluster could represent a novel type of mammalian gene cluster in which the juxtaposition of the genes is integral to the regulation of their expression.

Viral genomes tend to be very compact, possibly because of the need to economize on the amount of DNA which is necessary to be transferred from one host to another. The arrangement of viruses such as the papovaviruses is particularly reminiscent of the Surfeit cluster, indeed these viral genomes even contain multiple splice sites. If the Surfeit cluster is a "recent" addition then it must have integrated into the mammalian genome before the mammalian radiation about 80 myr ago because it is found in the same syntenic group in mouse and man which would be very unlikely to have happened by chance. Furthermore, the Surfeit cluster contains a number of repetitive B1 and B2 elements which conform well to the consensus sequences for these elements in mouse, indicating that it has been present long enough for these elements to integrate, though this may only be 3 myr or so (see section 1.2). The Surf-3 gene has been found to be an expressed gene for the ribosomal protein gene L7a (Giallongo et al 1989), this is a very ancient gene being highly conserved down to Yeast (J. Yon unpublished

results). As Surf-3 is located in the Surfeit cluster with other Surfeit genes on either side of it, this suggests that the Surfeit Cluster as a whole is also ancient and is indeed a cluster of mammalian "housekeeping" genes.

It is possible that the close spacing and overlapping of the Surfeit genes is simply a reflection of the enormous variability in the size and relative spacing of mammalian genes, and the observed flexibility of the genome when faced with transposable elements and dispersed highly repetitive DNA. Mammalian genes range in size from the 500 bp histone genes which have no introns (Marzluff & Graves 1984) to the Duchenne Muscular Dystrophy gene which consists of over 60 exons spread over 2000 kb (Burmeister et al 1988, Koenig et al 1987). The spacing of mammalian genes varies from the situation in α-satellite DNA where there are hundreds of kb of DNA but no known genes (Waye et al 1987), to the Tla region of the mouse major histocompatibility locus where 12 genes are located in 110 kb of DNA (Mellor 1986). Over 100,000 copies of the LINE element are thought to have integrated randomly into the mouse genome over the last three million years (Casavant et al 1988) and SINES, which occur at about 300,000 copies per genome, are often found in transcribed regions of the DNA where they do not appear to adversely affect gene expression.

If the DNA and RNA processing enzymes can handle such enormous variability and rapid change as described above, there is no reason why the Surfeit cluster should not represent a random clustering of housekeeping genes. Each gene would then be independently expressed and regulated, any affects of the close apposition of the genes, such as anti-sense interference, being incidental and insignificant. Examples of such close, or overlapping, arrangements of genes, which appear to have no functional significance are found in the Drosophila genome as described in the introduction (section 1.2). In Drosophila the *Dunce* gene which spanns 100 kb of genomic DNA contains at least three other genes in its large, 79 kb, intron. The internal

genes are transcribed in both the same and the reverse orientation as the Dunce gene allowing the possibility of both anti-sense inhibition and transcriptional interference of the internal promoters, to occur. However no functional significance of the cis organization of these genes has been found, the genes being effectively independently expressed (Chen et al 1987).

Alternatively, the juxtaposition of the genes could be related to the control of their expression by a variety of molecular mechanisms (section 1.1). These mechanisms have been demonstrated to affect eukaryotic gene expression in experimental and natural situations.

Bidirectional promoters which drive coregulated expression of the flanking genes have been demonstrated in Yeast. The two genes *Gal1* and *Gal10* are co-regulated by the trans acting transcription factor *GAL4* which binds between the divergent promoters (West et al 1984). On the other hand, bidirectional promoters which drive expression of one or other of the flanking genes but not both are observed in the papovaviruses. In SV40, the early genes are activated by ubiquitous cellular transcription factors such as SP1, but later during infection the early genes are repressed by the binding of the virally encoded large T antigen which also activates the transcription of the divergently transcribed late genes (Tooze 1981).

Interference of two promoters orientated in the same direction has been demonstrated to occur for two α-globin genes introduced into HELA cells on a single plasmid (Proudfoot 1986). Interference has also been shown to occur in vivo between the two promoters in the Long Terminal Repeats of retroviruses. The 5' promoter is only active if the 3' one is inactivated (Cullen et al 1984). Promoter interference is also thought to occur between the the two tandemly arranged acid phosphatase genes, *PHO5* and *PHO3* in Yeast. When the 5' *PHO5* gene is induced the 3' *PHO3* mRNA is no longer detectable. However in the absence of the *PHO5* gene no reduction of *PHO3* mRNA is seen under conditions which induce *PHO5* (Rogers et al 1982).

Antisense interference and down regulation of gene expression has been shown to occur experimentally with just 225 bp complementary to the 3' end of Thymidine Kinase mRNA. When a large excess of the anti-sense RNA is expressed in mouse TK+ L-cells, TK activity is reduced (Kim & Wold 1985). No examples of anti-sense inhibition being used naturally, as a means of gene regulation have been identified in eukaryotes. However, in a myelin deficient mutation of the mouse, *shi*^{mld}, it is possible that anti-sense inhibition is causing the loss of transcription of the myelin basic protein (MBP) gene. In this mutation, exons 3 to 7 of the MBP gene have been duplicated in an inverted orientation upstream of the complete MBP gene. Anti-sense RNA corresponding to these inverted exons is detectable and could be causing the repression of positive sense MBP mRNA (Okano et al 1988).

8.3 How to Resolve Chance from Function

There are two approaches to determining whether the close spacing of the Surfeit genes is a function of control of their expression, or merely a random occurrence which can be accommodated along with the enormous plasticity which is observed in the genome. The first is to continue with the characterization of the Surfeit genes, to identify their protein products and differential regulation, to see whether there is a correlation between the gene regulation and organization of the genes. The second is an experimental approach where one modifies the cloned Surfeit cluster DNA before reintroducing it into cultured cells to test whether various forms of *cis* interaction are occuring.

Only Surf-3 has been identified. For Surf-1, -2 and -4 the putative proteins encoded by the genes have been defined and it is now possible to generate antibodies against peptides or fusion proteins. One can then use the antibodies to characterize the distribution and function of the encoded proteins, with the aim of identifying them. Two new genes have been

identified in the Surfeit cluster and it will be interesting to clone cDNAs from these genes to try and identify them. The expression of the Surfeit genes has only been investigated in exponentially growing, tissue culture cells. Investigating the expression of the Surfeit gene under different physiological conditions, particularly those that affect ribosomal protein synthesis, and in different tissues may reveal differential regulation of these genes. The steady state levels of the mRNAs for Surf-3, Surf-2 and Surf-4 have been shown not to vary during the cell cycle, so this type of temporal regulation does not occur at the mRNA level though it might occur post transcriptionally.

An experimental approach to the functional importance of the central bidirectional promoter is being carried out by Dr. Andrew Lennard. The strategy is to place the bidirectional promoter between two marker genes and to transfect this, and mutated, constructs back into mouse cells. Binding factors in the bidirectional promoter have been identified by DNase footprinting and band shift assays. By introducing mutations into the factor binding sites, singly and in combinations, it is hoped to determine how the different factors affect transcription of the two flanking genes. At present these experiments have all been done with confluent Balb/C A31 cells where only a fraction of the factors may be acting so mutations in factor binding sites may appear to have no effect. These experiments can be extended to cells grown under different conditions, particularly conditions which are known to affect ribosomal protein gene expression, to see how this affects the functioning of the various transcription factors.

Now that the entire Surf-4 gene has been cloned as well as the entire Surf-2 gene, it is possible to make constructs to investigate the interaction of Surf-2 and Surf-4 experimentally. A construct can be made which contains both complete genes under the control of their own promoters but with the mRNAs marked by the addition of oligonucleotides. This allows the mRNA

transcribed from the introduced DNA to be distinguished from the endogenous mRNA. This construct can then be modified so that the complementary sequence at the 3' end of the mRNAs is removed, or so that there is a strong terminator of transcription between the two 3' ends, or so that one or other of the two promoters is inactivated. After transfection of these constructs into mouse cells, the relative levels of the two mRNAs can be measured to determine whether antisense inhibition or transcriptional interference is affecting the expression of these two genes. In addition, when antisense inhibition is occuring, the double stranded form of the interacting RNAs has been observed in the nucleus (Kim & Wold 1985, Yokoyama & Imamoto 1987). Detection of the double stranded RNA corresponding to the overlap of Surf-2 and Surf-4 in the nucleus by RNase protection would indicate that Surf-2 and Surf-4 are being expressed in the same cells and are interacting.

Similarly there is only 70 bp between the 3' end of the very highly transcribed Surf-3 gene and the 3' end of Surf-1. Because RNA polymerase II continues for some distance beyond the poly(A) addition site it is possible that the transcription of Surf-3 affects Surf-1. Both these genes lie in the 19 kb EcoRI fragment which could easily be cloned into a vector to analyse the effect of Surf-3 transcription on Surf-1. Particularly interesting would be to determine whether there is a strong stop to transcription in the 70 bp between the two poly(A) addition sites.

Promoter inhibition has been shown to affect promoters directing transcription in the same direction. This means that the Surf-3 promoter could be inhibiting the Surf-2 promoter. This can be investigated by making a construct containing the 19 kb EcoRI fragment, and similar constructs where the Surf-3 promoter has been inactivated. Transfection of these constructs into cells and measuring the relative expression of the genes will indicate whether promoter interference is occuring.

8.4 Conclusion

The Surfeit cluster consists of at least six, very tightly clustered genes. The 5' end of each gene is associated with a CpG-rich island and the genes are expressed in a variety of cell types indicating that they are housekeeping genes. One of the genes, Surf-3, has been identified as ribosomal protein L7a but the other Surfeit genes are probably not other ribosomal protein genes. There is no homology between the four characterised genes at the DNA sequence or protein sequence level indicating that the cluster has not arisen by gene duplication. The arrangement of six tightly clustered (and overlapping) genes is highly unusual in the mammalian genome and suggests that the organization is related to control of gene expression. Further characterization of the genes and gene products may reveal the identity of more of the Surfeit genes which will help in understanding the regulation of the genes. By manipulation of the cloned DNA it is hoped to be able to determine whether the genes are individually regulated or whether antisense regulation, promoter interference, or sharing of regulatory elements are being used as mechanisms of gene regulation in the mouse Surfeit gene cluster.

Chapter 9: Materials and Methods

9.1 Materials

Acrylamide and NN'-methylene-bisacrylamide	BDH. Stock solutions were deionised with"Amberlite" monobed resin MB-1 (BDH.) and filtered before use.
Agarose	FMC Bioproducts
Ampicillin	Penbritin; Beecham Research Laboratories
Bacto agar	Difco
Boric acid	FSA
BSA	Sigma
BSA enzyme grade	BRL
$CaCl_2$	BDH
CH_3COOH	FSA
CsCl	Koch-Light
$CHCl_3$	FSA
Chloramphenicol	Sigma
DNA 123bp ladder	BRL
Dyalysis Tubing	Spectrapor
DTT	Sigma
EDTA	FSA
ETOH	James Burough
Genescreen	NEN Research Products
Guanidine thiocyanate	Fluka (pure grade)
Glycerol	Fisons
Ficoll	Pharmacia
Films	Polaroid type 667 for photography of EtBr-stained agarose gels, Kodak XAR-5 and Fuji RX for autoradiography
Formaldehyde	Fisons (38% w/v solution)
Formamide	BDH (specially pure, no. 44254). Deionised using a mixed bed ion-exchange resin (Bio-Rad AG 501-X8)
Genescreen plus	NEN Research products
Glucose	BDH
HEPES	Sigma

Hexadeoxynucleotide primers	$pd(N)_6$ Pharmacia
H_3PO_4	BDH
8-Hydroxyquinoline	Sigma
IPTG	Sigma
Iso-amyl alcohol	Fisons
KCl	BDH
KH_2PO_4	BDH
KOAc	BDH
KOH	FSA
Lysozyme	Sigma
Maltose	BDH
MeOH	BDH
$MgSO_4.7H_2O$	BDH
$MnCl_2$	BDH
MOPS	Sigma
Na_2HPO_4	BDH
NaCl	BDH
NaOH	FSA
Nitrocellulose	Schleicher & Schuell
Nucleotides, deoxynucleotides, and dideoxynucleotides	Pharmacia
Oligo(dT)	Sigma
Oligo(dT)-Cellulose	Collaborative Research
PEG	Fisons
PIPES	Sigma
Radiochemicals	Isotopically labelled chemicals were obtained from Amersham International or NEN. Their specific activities are referred to in the text.
RbCl	Hopkin & Williams
RNA ladder	BRL
RNAsin	Human placental ribonuclease inhibitor Bolton Biologicals.
Salmon Sperm DNA	Sigma
Sarkosyl	BDH
SDS	BDH
Sephadex G75	Pharmacia

Sephadex G50	Pharmacia
Spermidine	Sigma
Sucrose	Fisons
Tetracycline	Sigma
Tris.HCl	Sigma
Triton X-100	Sigma
Tryptone	Difco
Urea	Fisons
Yeast extract	Difco
Yeast RNA	Sigma, deproteinized before use
X-Gal	Sigma
X-Ray intensifying screens	Du Pont Lightning Plus

EQUIPMENT

Laser Densitometer	LKB Ultroscan XL
Sequencing Apparatus	ATTO

All other reagents and equipment were obtained from Sigma or ICRF Central Services.

ENZYMES

Avian myeloblastosis virus RNA directed DNA polymerase (reverse transcriptase)	Anglian Biotechnology
Calf intestinal alkaline phosphatase	Boehringer
DNA restriction enzymes	New England Biolabs, BRL, Boehringer and NBL
DNase I	Worthington Biochemical Corporation
DNA polymerase I	Boehringer
E.coli DNA ligase	New England Biolabs
Klenow DNA polymerase	BRL
Lysozyme	Sigma
Nuclease S1	Sigma
Proteinase K	Sigma
RNase A	Sigma
RNase H	BRL
RNase T1	Sigma
Sp6 polymerase	Dupon

T4 DNA ligase	New England Biolabs
T4 DNA polymerase	P-L
T4 polynucleotide kinase	Pharmacia.

KITS

Gigapack gold	Stratagene
Lambda Dash	Stratagene
Lambda GT10	Vector Cloning Systems
Sequenase	United States Biochemical Corporation
Concatermerized Lambda markers	Clontech
GeneAmp	Cetus

Vectors and Host Bacteria

Lambda nm1149
Lambda GT10 (Stratagene)
Lambda Dash, BamHI ends (Stratagene)
Pxf3 (Hanahan 1983)
Bluescript (Strategene)
E.coli strain DH5
E.coli strain TG-1
XL1-blue recA-, (developed at Stratagene)

Cell Lines

TS-A-3T3	Balb/C 3T3 cells transformed by the polyoma virus TS-A mutant (Ruley et al 1982)
A31	3T3 cell line derived from Balb/c mouse embryo culture (Aaronson & Todaro 1968)
L1210	Leukemic cell line derived from a DBA female mouse (Moore et al 1966)
COS	Simian CV-1 cells transformed by an origin-defective mutant of SV40 which codes for wild-type antigen (Gluzman 1981)

Formulation of Frequently Used Solutions

GS1	2M NaCl 0.5M NaOH
GS2	4M NaCl

Pto.

	0.5M TrisHCl adjusted to pH 7.5 with NaOH
Tris saline	24mM Tris-base 0.8% w/v NaCl 0.38% w/v KCl 0.01% w/v Na_2HPO_4 0.1% w/v dextrose 0.001% w./ neutral red pH adjusted to 7.4 with HCl
PBS.A	1% w/v NaCl 0.025% w/v KCL 0.14% w/v Na_2HPO_4 0.025% w/v KH_2PO_4
PBS	10mM $CaCl_2$ 6mM $MgCl_2$ in PBS.A
Versene	0.54mM EDTA in PBS.A
TE	1mM EDTA 10mM Tris.Cl pH7.4
SM	10mM Tris.Cl pH7.4 10mM $MgCl_2$ 100mM NaCl 0.05% w/v gelatin
1xSSC	150mM NaCl 15mM sodium citrate pH to 7.5 with NaOH
1x TAE	40mM Tris Acetate 2mM EDTA
1x TBE	89mM Tris-base

pto.

89mM Boric acid
2mM EDTA

2FC
49.5% v/v redistilled phenol
49.5% v/v chloroform
1% v/v isoamyl alcohol
0.1% w/v 8-hydroxyquinoline.
saturated with TE pH 8.0

L Broth
1% w/v Bacto-tryptone
0.5% w/v Bacto-yeast extract
1% w/v NaCl
Supplemented with 1.5%w/v Bacto-agar for solid media
or, 0.6% agarose for an overlay for lambda

2xTY
1% w/v Bacto-tryptone
1%w/v Bacto-yeast extract
0.5% w/v NaCl
Supplemented with 1.5%w/v Bacto-agar for solid media

H top
1% w/v Bacto-tryptone
0.8% w/v NaCl
0.8% w/v Bacto-agar as an overlay for M13

9.2 Methods

DNA Quantitation and Manipulation.

DNA and RNA were quantitated by determining the absorbance of dilutions of stock solutions at a wave length of 260nm (1 OD_{260} = 50μg/ml DNA, 40μg/ml RNA). Following enzyme treatment, plasmid and cellular DNA was normally extracted with 2FC followed by $CHCl_3$ and ethanol precipitated before further manipulation. If small quantities of DNA were present 1-2μg of yeast tRNA, or glycogen where the absence of RNA was important, was added to assist ethanol precipitation. In addition, in the absence of any salt, either NaCl to 0.1M or NaOAc pH5.4 to 0.3M, was added to enable DNA precipitation.

Restriction enzyme digestions were performed as described in Maniatis et al (1982), except that 5 units of enzyme/μg of DNA was used. To digest

cellular DNA the buffers and enzymes were initially mixed in gently to prevent shearing. After 3 hr at 37°C, more restriction enzyme was added to the cellular DNA and the incubation continued for a further 2 hrs.

Intermolecular ligations were performed in small volumes, generally 10µl for a total DNA content of 100ng. For ligations involving plasmid sequences a 1:10 ratio of plasmid:insert DNA was used. When the plasmid was capable of self-ligation, the compatible ends were dephosphorylated before use (see below). Both vector and insert were often ligated directly from low melting temperature agarose (LMA) gels. To do this, the restriction enzyme digested DNA was run out on a LMA gel, the band containing the DNA was excised and placed in an eppendorf tube. Before adding to the ligation mixture, the LMA was melted by heating to 65°C for 5mins. LMA can be added to a final concentration of 0.4 % without having a deleterious effect on ligation. Ligations involving λ vectors were normally set up at three different ratios of arms:insert (1:2, 1:1 and 2:1 for the ratio of compatible arm ends : compatible insert ends) and after packaging and titration the most suitable ratio was selected for construction of a particular library. Ligations were performed overnight at 4-15°C in ligation buffer and proteins were not extracted before transfection into competent bacteria.

Dephosphorylation of DNA was performed using calf intestinal alkaline phosphatase as described in Maniatis et al (1982). Following incubation at 37°C the reaction was heated at 65°C for 10 min, and then extracted twice with 2FC.

ligation buffer	50mM Tris.Cl pH8.0
	10mM MgCl2
	1mM ATP
	10mM DTT
	50µg/ml BSA

Agarose Gel Electrophoresis of DNA.

Large submarine agarose gels (150-200ml) and mini submarine agarose gels (50ml) were made and run in 1xTBE or 1xTAE. Depending on the size of the DNA fragments under investigation, the agarose concentration varied from 0.7-2% w/v. All gels contained 0.5µg/ml ethidium bromide to enable visualization of the DNA under long wave (360nm) ultraviolet light. Samples were loaded following the addition of 0.1 x volume of loading buffer. To resolve fragments of cloned plasmids or lambda DNA, 1µg of DNA was digested

with the required restriction enzyme in 20µl of reaction mixture. Loading buffer was added directly after digestion with the restriction enzyme and both mini gels and large gels were run at 100mA.

To resolve fragments of genomic DNA from 0.5 to 30kb in size, 5µg of genomic DNA was digested in 400µl of restriction enzyme buffer with 20units of enzyme for five hours. The DNA was extracted with 2FC and $CHCl_3$ before precipitation with EtOH, drying and resuspending in 20µl of H2O. Loading buffer was added and the fragments resolved on a 0.8% large submarine gel run at 30mA overnight.

10 x loading buffer	25%w/v Ficoll
	50mM EDTA pH8
	0.1%w/v xylene cyanol
	0.1%w/v bromophenol blue

Isolation of Specific DNA Fragments from Gels.

This method was used for the isolation of DNA fragments from both agarose and acrylamide gels. The excised section of the gel containing the required band was placed inside a dialysis bag with TE plus 100µg/ml BSA. This was then placed in a gel tank of 0.5-1 x TBE and the DNA electroeluted at 100 volts for 1-2 hrs. The eluted DNA was extracted sequentially with 2FC and chloroform, ethanol precipitated, washed, dried and resuspended in TE.

Plasmid Vectors.

Plasmid vectors were either pxf3 (Hanahan 1983) or Buescript (Strategene). Recombinant plasmids were propagated on *E. coli* strain DH5 (see Maniatis et al 1982) or XL-1 blue (Stratagene), which were grown in L broth supplemented with 50 µg/ml ampicillin. In the absence of plasmids these bacteria were grown in L broth alone. For long term storage of bacteria, glycerol was added to a fresh overnight culture to 20%v/v, and an aliquot stored at -20°C.

Preparation, Storage and Transformation of Competent Bacteria.

(Hanahan 1985)

XL-1 blue or DH5 strains of *E. coli* were grown as isolated colonies on ψa plates. One colony was used to grow an overnight culture and 20 drops of this was used to inoculate 30ml of ψb medium. This was grown at 37°C with shaking for about 2hrs until the A_{550} was 0.3 when 25mls was added to 475ml

prewarmed ψb medium and grown for a further 2hrs at 37°C with shaking. When the A_{550} reached 4.8 the bacterial culture was placed on ice for 5mins and then harvested at 4°C. They were then resuspended in 200mls TFBI, kept on ice for 5mins, reharvested at 4°C and resuspended in 20mls of TFBII.

For long term storage, aliquots of 200µl were dispensed into eppendorf tubes using pre-chilled pipetts and tubes, and snap-frozen in dry ice. These bacteria were then stored at -70°C and remained fully competent for several years.

To transfect plasmid DNA into these competent bacteria, the frozen bacteria were thawed and up to 100ng DNA was added. The bacteria were kept on ice for 30-45mins then heat shocked at 42°C for 100secs and cooled on ice. 800µl ψb medium was added and the mixture incubated at 37°C for 45mins. They were then ready for plating out onto LB plates with suitable antibiotic selection. This method usually gives > 10^6 cfu/µg DNA

TFBI	30mM KOAc 100mM RbCl 10mM $CaCl_2$ 50mM $MnCl_2$ + glycerol to 15% pH to 5.8 with 0.2M CH_3COOH Filter sterilize
TFBII.	10mM MOPS or PIPES 75mM $CaCl_2$ 10mM RbCl + glycerol to 15% pH to 6.5 with KOH Filter sterilize
ψa.	5g/L Yeast extract 20g/L tryptone 5g/L $MgSO_4$ 14g/L Agar pH to 7.6 with KOH Autoclave
ψb	ψa without the agar.

Small Scale Preparation of Plasmid DNA

From an overnight culture of bacteria, 1.5mls were transferred to an eppendorf tube and spun for 3mins. The supernatant was aspirated off and the bacterial pellet respun for 1min and reaspirated. The pellet was resuspended in 100ul STET (see below) with 2mg/ml lysozyme added fresh. Each sample was stored on ice until they had all been resuspended and they were then all placed in a boiling water bath for 45 seconds. After centrifuging for 15mins the supernatant was taken and added to an equal volume of isopropanol, placed at -20°C for 10mins and respun for 10mins. The resulting DNA pellet was washed with 70% EtOH and resuspended in TE with 0.1mg/ml RNase A. For bluescript vectors this yields about 30µg of DNA.

STET
50mM Tris.Cl pH 8.0
50mM EDTA pH 8.0
8% w/v sucrose
5% v/v Triton x-100

Large Scale Plasmid Preparation

(Holmes & Quigly 1981 with variations)

A 10ml overnight bacterial culture was used to inoculate 400mls of prewarmed LB media in a 2L flask and this was incubated, with shaking, at 37°C. When the OD_{600} reached 0.4, chloramphenicol was added to 150µg/ml and the culture continued overnight. For bluescript vectors, no amplification is needed and chloramphenicol was not added. The bacteria were pelleted by centrifugation at 10,000g for 7mins, washed in bugwash and repelleted. The cells were then resuspended in 10mls of STET (see above) containing 4mg/ml of lysozyme added fresh, and immediately placed in a boiling water bath for 2 mins. The resultant glop was transferred to SW41 tubes and centrifuged at 30k for 30mins at 4°C in an SW41 rotor. The supernatant from this centrifugation was taken and 1g CsCl added per ml and 0.4ml of 10mg/ml EtBr added per 10mls. This solution was transferred to 2 vti65 tubes, the tubes were sealed and centrifuged in a vti65 rotor at 64,000rpm for 5hrs at 18°C.

After centrifugation the bands of plasmid DNA can be easily seen due to the ethidium bromide and the lower, super coiled, band removed using a 23gauge needle. The EtBr was removed by repeated extraction with CsCl saturated isopropanol until no red colour could be seen. The salt

concentration was reduced by the addition of twice the volume of water and the DNA precipitated by the addition of 2.5 times the volume of 95% EtOH. The DNA was pelleted by spinning at 10,000g in a sorval rotor, washed with 70% EtOH, dried and then dissolved in TE buffer. For bluescript vectors a 400ml culture yields about 4mg of DNA.

Bugwash	0.15M NaCl
	10mM Tris.Cl pH 8.0
	1mM EDTA

Oligonucleotides

The oligonucleotides used for primer extension and Polymerase Chain reaction, were obtained from Dr. Iain Goldsmith of the ICRF. These oligonucleotides have been fully deprotected, have -OH groups at both the 3' and 5' ends and the stepwise efficiency was 98.6%. After dissolving in TE the concentration was calculated from the UV absorbance at 260nm, assuming that a solution containing 33mg/ml has an absorbance of 1.0.

Polymerase Chain Reaction

The polymerase chain reaction (PCR) was used to amplify and clone fragments from the published sequences of ribosomal proteins L32 and L30. GeneAmp DNA Amplification Reagent Kit purchased from Perkin Elmer Cetus was used for most of the reagents. The reaction mixture contained 2µg of LTK- mouse cellular DNA, 200µM each dNTP, 1.0µM each primer, 2.5 units Taq polymerase, 50mM KCl, 10mM Tris-Cl (pH 8.3), 1.5mM $MgCl_2$ 0.1% (w/v) gelatin, in 100µl. This was placed in a small microfuge tube and covered with 100µl of mineral oil to prevent evaporation. The PCR reaction was carried out overnight on a Programmable Dri-Block purchased from Techne. 30 cycles of 2mins at 94°C, 2mins at 59°C, and 9.9min at 70°C were followed by an additional 9.9mins at 70°C. The PCR reaction products can be stored at 4°C for several weeks.

To subclone the PCR product, one tenth of the PCR reaction was kinased with polynucleotide kinase and ATP and this DNA size fractionated on a low gelling temperature agarose gel. The products normally appeared as a clearly distinguishable band which was excised and ligated directly into a bluescript vector cut with SmaI and phosphatased.

Mammalian Cell Culture

All mammalian cell lines were grown in a 10% CO_2: 90% air humidified incubator, using Dulbecco-modified Eagles medium (DMEM) supplemented with 0.1% w/v glutamine and 5% v/v FCS. Cells were grown on plastic tissue culture dishes (9cm, circular) or 24.5 x 24.5 cm bioassay trays (Nunc). Cells were removed from the dishes, following an initial versene wash, with a 3:1 mixture of versene : 0.1% w/v trypsin in Tris saline, and passaged at appropriate densities.

Preparation of High Molecular Weight Genomic DNA.

Eukaryotic cell DNA was prepared by a variation of the procedure of Blattner et al (1978). Cells were cultivated on large bioassay trays until almost confluent. After removal of the medium, cells were washed twice in PBS and drained thoroughly. Each tray of cells was then overlaid with 4ml of lysis buffer and left at room temperature for 10 mins. The lysed mixture was then carefully scraped into a 50ml Falcon tube, using a rubber policeman. The lysed cells were then extracted twice with 2FC and once with chloroform. The DNA was subsequently dialysed against 2L 1M NaCl, TE for 12 hrs at 4°C and then 4L TE for 48 hrs at 4°C. The nucleic acid was then treated with 20µg/ml RNase A at 37°C for 3 hrs, protein extracted as above and dialysed against 4L TE for 48 hrs at 4°C before storage at 4°C.

lysis buffer 10mM TrisHcl pH8
10mM EDTA pH8
5mM NaCl
0.5% w/v SDS
0.1mg/ml proteinase K

Size Fractionation of Restricted Total Cellular DNA.

In order to enrich total enzyme restricted cellular DNA for a particular fragment, the mixture of fragments was size fractionated on a sucrose gradient. Restriction enzyme digests of total cellular DNA (500µg) were loaded onto 10-40% sucrose gradients and spun at 120,000g (24,000rpm)/10°C/20 hrs/SW27. Fractions of 500µl were collected via a 19 gauge needle inserted through the base of the tube. In order to determine the average size of DNA fragments in each fraction, 50µl aliquots of every third fraction were run on a 0.8% mini agarose gel, using λ DNA digested with HindIII as a size marker. Having determined roughly which fraction should contain the fragment of interest, 200µl aliquots of each fraction in the

appropriate range were precipitated, run out on 0.8% agarose gels and blotted onto genescreen plus as described under southern blotting. The fraction(s) containing the desired sequences were then identified by hybridization to radiolabelled DNA probes.

Sucrose soln. 5mM EDTA
20mM Tris.Cl pH 8.0
1M NaCl
10% or 40% Sucrose

Preparation of High Molecular Weight Genomic DNA in Agarose Blocks

Genomic DNA as prepared by the method described above contains fragments of approximately 100kb in size due to shearing of the DNA and to the presence of trace amounts of DNase. To prepare, and manipulate, DNA of greater than 100kb in size, the cells from which the DNA was to be obtained were first immobilized in low melting agarose blocks and all steps were carried out so as to exclude any contamination by DNase.

DNA was prepared from tissue culture A31 cells or directly from the spleen of DBA mice. Tissue culture cells were grown to near confluency and then trypsinised to give a single cell suspension, whereas the whole spleen was cut into small chunks and then homogenised in PBS, again to give a single cell suspension. The single cell suspensions were washed in PBS and the concentration of cells measured with a coulter counter, they were then resuspended in PBS at $1x10^6$ cells per 40µl PBS at room temperature, $1x10^6$ cells is equivalent to 10µg of genomic DNA. A 1% solution of low melting agarose in PBS was melted and held at 42°C, an equal volume of this was mixed with an equal volume of the cell suspension at room temperature. This mixture was dispensed immediately into 80µl block formers and allowed to set on ice.

To prepare the DNA in the agarose blocks, the blocks were incubated in 50mls of 1% Sarkosyl, 0.5M EDTA pH8, and 2mg/ml Proteinase K for 48hrs at 55°C. The blocks were then rinsed three times in TE buffer and placed in TE plus 0.04mg/ml PMSF for 30 minutes at 55°C, this was repeated once. The blocks were stored in 0.5M EDTA pH8 at 4°C.

Before using the blocks, they were rinsed three times for 1/2hr in TE buffer. To cut the DNA with restriction enzymes, a single block was placed in an eppendorf tube with 400µl of restriction enzyme buffer recommended by the manufacturer, and equilibrated at 37°C for an hour. The buffer was then

replaced and 50 units of enzyme added, digestion was carried out at the recommended temperature for that particular enzyme over night. New buffer and enzyme was added next day and the digestion continued for 2 hours. Before loading onto the gel the blocks were rinsed twice in TE buffer.

For size markers greater than 50kb in size, ligated lambda obtained from Clontech was used.

Pulsed Field Gel Electrophoresis:

To separate DNA fragments greater than 30kb in size, Pulsed Gel Electrophoresis was carried out using an LKB Pulsaphore unit with a hexagonal adaptor. A 1% agarose gel in 0.25x TBE was poured using the gel plate and former supplied by LKB. The agarose blocks were inserted into the gel slots and sealed in place with 1% low melting agarose. The gel was run at 170V with a pulse time of 30 seconds for 15hrs to resolve DNA up to 500kb in size, or at 170V with a pulse time of 100 secs for 48hrs to resolve DNA up to 3,000kb. The gel was placed in a solution of EtBr for half an hour and the DNA visualized with UV light. The gel was then used directly for southern blotting.

Southern Blotting:

(Church and Gilbert 1984)

To prepare blots of fragments less than 30kb in size, 5µg genomic DNA was digested with the appropriate restriction enzyme, phenol/chloroform extracted, precipitated with EtOH and loaded onto a 0.8% agarose gel. Both gel and running buffer contained 1x TBE, and the gel contained EtBr so that the DNA could be visualized and photographed with UV light after running the gel. For larger fragments a pulsed field gel was run as described above.

The DNA in the agarose gel was depurinated by soaking the gel in 0.25M HCl for 20mins on a shaking platform. The DNA was then denatured by soaking the gel in 0.5M NaOH /1.5M NaCl for 2x15mins, again on a shaking platform. Genescreen membrane was put onto H_2O to moisten and then rinsed in the denaturation solution above. DNA was transferred from the gel to the Genescreen membrane by blotting using a reservoir of denaturation solution and a stack of paper towels.

Transfer was allowed to continue overnight and then the filter neutralized by submerging in 50mM NaPi, pH 6.5 for 10-20 mins. After being allowed to dry on the bench, the membrane was baked at 80°C for 2hrs and

then exposed to 254nm UV radiation for 2mins from a distance of 50cm (approximately $0.16KJ/m^2$)

Hybridization of Southern and Northern blots

To hybridize, the filter was placed in a sealed plastic bag with 30-40ml of hybridization solution and prehybridized for 5min at 65-68°C. The prehybridization solution was removed and replaced with 10ml of new hybridization solution containing denatured probe. Radioactive labelled probe was used at a concentration of <10ng/ml or $<1.5x10^6$ cerenkov counts/ml, it was prepared by oligo labelling (see below) and was denatured by the addition of NaOH to 0.4M for 5mins.

Hybridization was allowed to continue overnight and the filter then washed in a solution of 40mM NaPi buffer, 1% SDS at 68°C. Washing was carried out for 3x5mins and then 30mins. The filter was then placed on the plastic side of a piece of benchcoat, wrapped in Saran Wrap, and exposed at -70°C with an intensifying screen for 1-7days.

Filters were stripped by washing for 30min at 80°C in 2mM EDTA, 0.1% SDS after which they did not need prehybridization.

1M NaPi buffer:	89g Na_2HPO4 x$2H_2O$ adjust to pH 7.2 with conc. H_3PO_4
Hybridization soln.:	0.5M NaPi pH7.2 7% SDS 1mM EDTA

Oligo-labelling of DNA Fragments by Hexadeoxynucleotide Primers.

Preparation of DNA: Plasmid DNA was cleaved with appropriate restriction enzyme(s) to give at least 500ng of the required fragment(s) for labelling. The digest was run out on a 1% low gelling temperature agarose gel (BRL 5517) and the required band excised after visualizing with UV light. Water was added to the agarose block containing the DNA to give a final concentration of 50ng DNA/20µl, this was put in a boiling water bath for 5mins and then kept at 37°C for 10-60mins prior to use. Remaining DNA was stored at -20°C and simply reboiled and stored at 37°C before use.

Labelling: The labelling reaction was carried out at room temp. by addition of the following reagents in the stated order;

5µl OLB buffer

Pto.

1µl of 10mg/ml BSA (enzyme grade, BRL 5561)

15µl DNA solution prepared as above or containing 25-50ng denatured DNA.

2.5µl ^{32}P α-dCTP (10mCi/ml, 3000Ci/mmol)

2.5µl ^{32}P α-dATP (10mCi/ml, 3000Ci/mmol)

1µl (4u) Klenow (large fragment) Polymerase 1.

OLB buffer is made by mixing solns. A:B:C in the ratio 10:25:15

Solution A 1000µl of 1.25M Tris-HCL, .125M $MgCl_2$.

18µl 2-mercaptoethanol

5µl of 0.1M dGTP, and 0.1M dTTP

Solution B. 2M HEPES (pH to 6.6 with 4M NaOH)

Solution C. Hexadeoxynucleotides, suspended in TE at 90 OD units/ml

The reaction was incubated at room temp for at least 5hrs or overnight, diluted with 150µl of NTE and heated to 67°C for 3mins to melt the LM agarose. The labelled DNA was separated from the unincorporated nucleotides by passing the reaction mixture with NTE over a G75 column and collecting the peak of radioactivity as it came off the column. This method yields 25ng DNA labelled to 10^8dpm/mgDNA, or a total of 2.5 x10^6 dpm.

Isolation of Mammalian Total Cellular RNA and Poly (A)+ Selection.

(Chirgwin et al 1979)

Nearly confluent cells were washed twice with PBS and thoroughly drained. The cells were lysed at RT with 4M GT (50% w/v guanidinium isothiocyanate, 0.5% v/v Sarkosyl, 25mM HEPES, 100mM β-mercaptoethanol, pH adjusted to 7 with NaOH and filtered) using 5ml per bioassay dish . High molecular weight DNA was sheared by vortexing, and passage of the extract through a 21 gauge needle. The extract was then layered onto an equal volume cushion of 5.7M CsCl, 25 mM NaOAc pH 5.4, and centrifuged at 200,000g (35,000 rpm)/20°C/20-24 hrs in a Beckman SW40 rotor.

After centrifugation the supernatant was discarded and the tubes were sliced through just above the RNA pellet to prevent contamination from adherent protein or DNA. The pellets were transferred to new tubes and washed with 70% v/v ethanol. The RNA was redissolved in 10mM Tris.Cl pH7.4, 5mM EDTA, 1% w/v SDS and extracted with an equal volume of 80% v/v chloroform, 20% v/v butan-1-ol. The organic phase was re-extracted, the two aqueous phases combined and NaOAc pH5.4 added to 0.3M. RNA was then ethanol precipitated, resuspended in 0.3M NaOAc pH5.4, transferred to an

Eppendorf tube and reprecipitated. The RNA was either stored in water, or under ethanol, at -20°C.

RNA was fractionated into poly(A)+ and poly(A)- components by chromatography on oligo(dT)-cellulose as described by Maniatis et al (1982). The yield of poly (A)+ RNA varied from 2-5% of the total RNA. When the amount of RNA present was low, ethanol precipitation of the RNA from 0.3M NaOAc pH 5.4 was assisted by the addition of 1-2µg of RNase free dextran.

Formaldehyde Agarose Gel Electrophoresis of RNA and Northern Blotting.

The procedures used were modified from those presented by Maniatis et al (1982). 200ml gels were cast which contained 1% w/v agarose in 1xMOPSE, 18% v/v formaldehyde. Samples for electrophoresis, usually 1-2µg of poly(A)+ RNA or 20µg of poly(A)- RNA, were prepared in 1xMOPSE, 6.8% w/v formaldehyde, 50% v/v formamide and heated at 55°C for 15-30 mins, following which 0.1x volume of RNA loading buffer (50% v/v glycerol, 1mM EDTA, 0.4% bromophenol blue, 0.4% xylene cyanol) was added per sample prior to loading on the gel. Gels were electrophoresed in 1xMOPSE, 6.8% w/v formaldehyde at 30mA overnight with the buffer circulating by means of a peristaltic pump.

Before blotting, the gels were soaked in water for 5 mins, followed by 50mM NaOH for 20-30 mins, to partially hydrolyse the RNA for ease of transfer during blotting. Gels were then neutralised by soaking in 0.1M Tris.Cl pH7.4 for 20-30 mins and finally soaked in 20xSSC for 60 mins. The gel was laid on a watman filter paper wick which dipped into 20xSSC, nitrocellulose or genescreen plus placed on top and then a thick wad of kleenex paper towells added with a weight (The nitrocellulose was first pre-wet in water, and then placed in 20xSSC for 5 min, before being applied to the gel). After 3-12 hrs of RNA transfer, the nitrocellulose was air dried, without rinsing, and then baked under vacuum at 80°C for 2 hrs.

Hybridization and washing procedures were identical to those used for Southern blotting. RNA markers (BRL) were run in a separate lane to one side of the gel, these were blotted in the same way as the rest of the gel and then stained to visualize the RNA bands. The filter was soaked in 5% v/v acetic acid for 15 mins and then in 0.5M NaOAc pH5.4, 0.04% w/v methylene blue for 5-10 mins. The filter was then rinsed with water to reveal the RNA.

1XMOPSE 20mM MOPS
1mM EDTA

Pto.

5mM NaOAc

pH adjusted to 7.0 with glacial acetic acid

cDNA Synthesis.

Oligo(dT)-primed first strand synthesis was initiated by adding 10µl oligo(dT) (1mg/ml), 5µl Tris-HCl (1M, pH 8.3 at 42°C), 5µl $MgCl_2$ (0.2M) and 14µl KCl (1M) to 5 µg TS-A 3T3 poly (A)+ RNA in 27.5µl. After heating to 65°C for 10mins and briefly chilling on ice, the following were added: 1µl DTT (1M), 2.5µl RNAsin, 2µl (α-^{32}P)dCTP (3000 Ci/mMol), 4µl each dCTP, dATP, dGTP and dTTP (25mM), and 5µl super RT (AMV reverse transcriptase).

Reverse transcription was performed for 2 hours at 42°C, after which 25µl of 0.1M EDTA was added. Labelling the cDNA to a relatively low specific activity in this way provided a useful means of detection of small quantities of product at later stages. A small aliquot of the reverse-transcribed product was run out on a sequencing-type gel in order to assess its average size. The solution was extracted with 2FC followed by $CHCl_3$, and the nucleic acid precipitated with ethanol following the addition of NH_4OAc to 2M. The samples were resuspended in 2M NH_4OAc, reprecipitated and washed with 70% v/v ethanol before further use.

Second strand synthesis was performed by the method of Gubler and Hoffman (1983), in SSS buffer (20mM Tris.Cl pH 7.4, 5mM $MgCl_2$, 10mM $(NH4)2SO_4$, 100mM KCl, 0.15mM β-NAD, 40mM each dNTP, 50µg/ml BSA) using 6 units/ml E.coli RNase H, 100 units/ml DNA polymerase I and 80 units/ml E.coli ligase. Reactions were incubated at 12°C for 60 mins, followed by 22°C for 60 mins. Reactions were stopped by adding EDTA to 20mM and SDS to 0.1% w/v final. Samples were extracted twice with 2FC, once with chloroform and ethanol precipitated twice in the presence of 2M NH_4OAc.

To give the ds cDNA blunt ends, the ds cDNA was resuspended in T4 DP buffer (33mM Tris-acetate pH 7.9, 66mM KOAc, 10mM MgOAc) and treated with T4 DNA polymerase for 5 mins at 37°C . Subsequently dNTPs were added to a final concentration of 200mM and the incubation continued at 37°C for a further 30 mins. Samples were extracted with 2FC, then chloroform and ethanol precipitated in the presence of 2M NH_4OAc.

EcoRI linkers were then ligated to the repaired cDNA, and excess linkers removed by digestion with EcoRI. The ds cDNA was purified away from single linkers by agarose gel electrophoresis and size-selected to be greater than 300 bp in length. The cDNA was then cloned into the EcoRI site of λgt10 and screened as described below.

Bacteriophage λ Vectors.

To make the lambda library of size selected BamHI fragments, the λ replacement vector EMBL4 (Frischauf et al 1983) was used. The lambda DNA was restricted with BamHI and SalI and then precipitated with 0.2M NaCl and an equal volume of isopropanol. This procedure does not precipitate the small BamHI to SalI fragments, so the 14kb stuffer region cannot ligate to the arms. 1μg of insert DNA was ligated to 1μg of lambda DNA and the mixture packaged with Gigapack Gold purchased from Stratagene. Small aliquots of the library were plated out on the E. coli strains WL95 (a P2 lysogen on which only recombinant EMBL4 gives lysis) and WL87 (both recombinant and nonrecombinant EMBL4 gives lysis). This showed that the library contained nearly 100% recombinants out of a total of 100,000. The whole library was then plated out on E. coli WL87, as this is rec- whereas WL95 is rec+.

For the library of size selected BclI fragments, lambda Dash cut with BamHI, purchased from Stratagene, was used. 1μg of insert DNA was ligated to 1μg of vector and the mixture packaged with Gigapack Gold purchased from Stratagene. Aliquots were plated out on E. coli strains P2392 (only recombinant phage give lysis) and LE392 (both recombinant and non-recombinant phage give lysis) showing that the library contained 90% recombinants out of a total of 150,000. The whole library was plated out on E. coli strain LE392 for screening.

For the cDNA library the lambda insertion vector λNM1149, which is very similar to λGT10, was used EcoRI digested lambda DNA was ligated with cDNA (with EcoRI compatible ends) and packaged with Gigapack, purchased from Stratagene. Small aliquots of the library were plated on E. coli C600 hfl (this strain contains lambda repressor and lysis only occurs with recombinant NM1149) and strain WL87 (both recombinant and wild type NM1149 give lysis) showing that the library contained 10% recombinants out of a total library of 300,000. The whole library was plated out on E.coli C600 hfl as this is rec-, for screening.

The library of non-size selected EcoRI fragments was cloned into λGT10 purchased from Stratagene, 0.2 μg of insert DNA was ligated to 1 μg of λGT10 and packaged with Gigapack gold using the manufacturers recommended conditions. Small aliquots were grown on E. coli strain C600 hfl (on these baceria only recombinant phage give lysis) and strain C600 (on this strain both recombinant and nonrecombinant GT10 give lysis) showing that the

library was 45% recombinants out of a total of 110,000. The whole library was plated out on E. coli C600 for screening.

Bacteriophage λ Libraries: Plating out and Screening.

Lambda libraries were packaged using kits purchased from Stratagene using the protocols provided. To make competent bacteria, a 50ml overnight culture of the appropriate bacteria was spun down and resuspended in 20ml of ice cold 10mM $MgSO_4$. The required fraction of the packaging mix was added to 1ml of SM plus 1ml of the bacteria in a plastic universal and incubated at 37°C for 15 mins without shaking. The bacteria were mixed with 20mls of molten top agarose at 42° and quickly poured onto a bioassay tray containing 400ml of solid L broth agar supplemented with 10mM $MgSO_4$ and 0.2% Maltose, pre-warmed to 37°C. Libraries were plated out on large bioassay trays to give about 50,000 plaques per plate, although a maximum of half of any particular packaging reaction was used per plate. Once set, the plates were inverted and incubated for 12 hrs at 37°C. Plates were then transferred to 4°C for 2 hrs, to allow the top agarose to harden, prior to the taking of nitrocellulose lifts.

Nitrocellulose lifts of phage DNA were taken in duplicate to guard against false positives, and the filters and plates marked with ink to facilitate later alignment. Nitrocellulose was placed on the plates for 30 secs for the first lift, and for 60 secs for subsequent lifts. Subsequently the nitrocellulose was placed in a bath of GS1 for 30 secs, GS2 for 2-5 mins and finally rinsed in 2xSSC before being air dried and baked under vacuum at 80°C for 2 hrs. Conditions for hybridization and washing of the filters were the same as for Southern blots. Positive plaques were picked and purified to homogeneity by replating out and screening before further use.

Small Scale Preparation of Bacteriophage λ DNA

A 10ml overnight culture of the appropriate bacteria was spun down and resuspended in 4ml of ice cold 10mM $MgSO_4$.(These bacteria can then be stored at 4°C for up to 2 weeks). A single well isolated λ plaque was removed from an agar plate in a plug of agar and this was added directly to 5mls of a 1:100 dilution of the above bacteria in LB broth supplemented with 10mM $MgSO_4$ and 0.2% Maltose. The culture was incubated overnight at 37°C and λ DNA was isolated from the resulting liquid lysate as described by Maniatis et al (1982). This produces about 5μg of λ DNA.

Large Scale Preparation of λ DNA.

A 10ml overnight culture of the appropriate bacteria was spun down and resuspended in 4ml of ice cold 10mM $MgSO_4$. These were then incubated with approximately $5x10^7$ phage for 15 min at 37°C and then used to inoculate 400ml of prewarmed L broth supplemented with 20mM $MgSO_4$ which was then shaken vigorously overnight. Lysis of the bacteria is indicated by a granular appearance of the culture, rather than the normal silky appearance of *E. coli*. The purification of phage DNA from the liquid lysate by centrifugation through CsCl step gradients was as described by Maniatis et al (1982).

End-labelling DNA.

DNA for 5' end-labelling was dephosphorylated and free of contaminating RNA. 1μg of the DNA was labelled with 100μCi (γ-^{32}P)ATP (>5,000 Ci/mmol) using T4 polynucleotide kinase in kinase buffer (50mM TrisHCl pH7.5, 0.1mM EDTA, 10mM $MgCl_2$, 5mM DTT, 0.1mM spermidine) for 60 mins at 37°C. The volume was then increased to 200μl, containing 0.3M NaOAc pH 5.4, and the mixture extracted with 2FC, then chloroform and finally ethanol precipitated. Where the labelled DNA was to be used for size markers no further manipulation was necessary. However when the labelled DNA was to be used for sequencing, S1 mapping, or primer extension, the labelled end required had first to be purified away from the other labelled end. In most cases this was performed by cleaving with a second restriction enzyme that produced fragments that could be easily separated on non-denaturing acrylamide preparative gels (see below). If no convenient second enzyme existed, or when single stranded sequences were required for primer extension analysis, the strands were separated on acrylamide "cracking" gels (see below). In order to ascertain which was the correct strand to be used, both strands from "cracking" gels were Maxam-Gilbert sequenced before use.

In order to label recessed 3' termini, plasmid digests or purified DNA fragments (0.2 to 1 μg) were incubated in 30μl of restriction enzyme buffer containing 20μCi of the appropriate α-labelled dNTP (3000 Ci/mMol) and 1 unit of Klenow polymerase for 15 mins at room temperature. The volume was then increased to 200μl, containing 0.3M NaOAc pH 5.4, and the mixture extracted with 2FC, then chloroform, and finally ethanol precipitated.

Single Stranded RNA Probes.

The DNA fragment corresponding to the required RNA probe was cloned in the appropriate orientation into the vectors pSP64 or pSP65. CsCl purified

plasmid DNA was linearized, extracted with 2FC and $CHCl_3$, washed with 70% ETOH,dried and made up in TE at 1mg/ml.

The reagents for the labelling reaction are added in the order given,

4.0µl	5x Transcription buffer
2.0µl	100mM DTT
0.8µl	25u/ml RNasin
4.0µl	2.5mM ATP/CTP/UTP
2.4µl	100mM GTP unlabelled
1.0µl	0.2-1µg Linearized DNA template
5.0µl	50mCi ^{32}P-GTP (>400Ci/mM) 10mCi/ml
1.0µl	5-15u SP6 RNA Polymerase.

5xTranscription Buffer:	30mM $MgCl_2$
	200mM Tris-HCL pH 7.5
	10mM Spermidine
	50mM NaCl

The reaction mixture is incubated for 60mins at 37°C, then RNase-free DNase was added to a concentration of 2u/µg of DNA and the incubation continued for 15mins at 37°C.

If the labelled RNA was to be used for probing Northern blots then it was separated from the unincorporated nucleotides by passing over a G-50 spun column as described in Maniatis et al (1982). If it was for RNase protection then it was gel purified.

End Labelling Oligonucleotides

End labelling was carried out using the following reaction mixture:

1µl 10ng/ml oligonucleotide
1µl 10x Buffer
1µl 0.05M DTT
2.5µl ^{32}PγATP
3.5µl H_2O
1µl PNK

The reaction was carried out at 37°C for 30-60mins, labelled oligonucleotides were separated from unincorporated nucleotides by passing through a G-50 spun column as described in Maniatis et al (1982), or by running on a 20% denaturing acrylamide gel.

Acrylamide Gel Electrophoresis.

All acrylamide gels were polymerised using 250µl 10% w/v ammonium persulphate and 30µl of TEMED per 50ml gel volume, and pre-run before use. Denaturing acrylamide gels were used for sequencing, primer extension product analysis, S1 product analysis, isolation of full length end labelled oligonucleotides and full length single stranded RNA probes. These gels were 0.4mm thick, contained from 4-20% polyacrylamide (using a stock solution of 30:1 acrylamide:bisacrylamide) in 1xTBE buffer with 50% w/v urea. Electrophoresis was in 1xTBE at a current sufficient to maintain denaturing conditions within the gel (normally about 45W). Samples obtained from dideoxy sequencing reactions were treated as outlined in Bankier and Barrell (1983) before loading. All other samples were resuspended in form stop buffer and heated to 90-100°C for 2 mins before loading.

For autoradiography of the resolved fragments, preparative gels containing ^{32}P radiochemicals were normally covered in Saran wrap and autoradiographed at RT. All gels containing ^{35}S radiochemicals were fixed for 30 mins in a tray of 2L of 10% v/v methanol, 10% v/v acetic acid, and then dried onto 3MM Watman filter paper at 80°C for 30mins. Autoradiography without a screen was carried out at RT over night.

Strand separating ("cracking") gels allow the two strands of a fragment of double stranded DNA fragment to migrate at different speeds due to the formation of different secondary structures. These gels were 0.8mm thick and were typically 5% polyacrylamide in 0.5xTBE buffer using a stock solution of 66:1 acrylamide:bisacrylamide. Samples were denatured in form stop buffer, or strand separation buffer (50% v/v dimethyl sulphoxide, 1mM EDTA, 0.02% w/v bromophenol blue and xylene cyanol), by heating to 90-100°C for a few minutes before loading. The gels were run cold (300V) in 0.5xTBE overnight. Generally end labelled fragments were used, and if both ends were labelled, the individual strands were identified by Maxam and Gilbert sequencing.

Formamide stop	10mM NaOH
	1mM EDTA
	80% v/v formamide
	0.02% w/v bromophenol blue
	0.02% w/v xylene cyanol

S1 Nuclease Mapping.

(Weaver and Weissman 1979)

A typical reaction contained 20,000 cpm (Cerenkov) of probe DNA (approximately 5ng) mixed with either 40µg of total RNA or 2-5µg of poly (A)+ RNA, whilst negative control reactions contained 40µg of yeast RNA. After mixing the RNA and DNA were precipitated with EtOH, dried and the sample resuspended in 20µl of annealing buffer. Samples were denatured at 85-90°C for 15 mins, and then immediately transferred to a water bath, set at the required temperature, for overnight annealing. The temperature was set to maximise RNA:DNA hybridization (as opposed to reannealing of double stranded DNA probes) from the equation determined by Thomas et al. (1976):

tmax = 81.5 + 0.5(%G+C) + 16.6 (log[Na+]) - 0.6(%formamide)

which, for the annealing buffer used, can be simplified to:

tmax = 26.9 + 0.5(%G+C)

Thus the temperature of hybridization varied between different experiments, depending on the (%G+C) content of the probe. After overnight hybridization, without allowing any prior cooling of the annealing mix, 300µl of ice cold S1 digestion buffer was added to each sample, which was quickly mixed and placed on ice. When all the samples had been treated in this way, they were incubated at 37°C for 30 mins, or at 15°C for 1-2 hrs when shorter protected fragments were expected. Reactions were stopped by the addition of 75µl of S1 stop buffer, 2µg of yeast RNA added and the samples precipitated on dry ice using 400µl of isopropanol. Washed, dried samples were then resuspended in 5µl of form stop buffer, heated to 90-100°C for 5 mins and loaded onto acrylamide sequencing gels.

annealing buffer	40mM PIPES pH 6.4
	1mM EDTA pH8.0
	0.4M NaCl
	80% v/v formamide
s1 digestion buffer	280mM NaCl
	50mM NaOAc pH 4.6
	4.5mM $ZnSO_4$
	20µg/ml denatured salmon sperm DNA
	400units/ml nuclease s1
s1 stop buffer	50mM EDTA
	2.5M NH_4OAc

Primer extension

Oligonucleotides complementary to the mRNA under investigation were end labelled as described above. 40,000 Cerenkov counts of end labelled oligonucleotide was mixed with 1 to 5 µg of Poly(A)+ RNA and 40µg of tRNA, or 40µg of tRNA alone. The mixture was precipitated with 2M NH_4OAc and EtOH, washed with 70% ETOH, dried and resuspended in 20µl of annealing buffer as for S1 mapping. The mixture was heated to 90°C for 5 mins and then annealed over night at 30°C for Surf-3 or 45°C for Surf-4.

After overnight hybridization, without allowing any prior cooling of the annealing mix, 400µl of ice cold 2M NH_4OAc was added and the nucleic acid precipitated on dry ice using 1ml of EtOH. Primer extension reactions were performed using a modification of the method of Huynh et al (1985). Washed, dried samples were resuspended in 19µl of First Strand buffer and 25 units of reverse transcriptase added. Samples were then incubated at 42°C for 60 mins and the reactions stopped by the addition of EDTA to 20mM.

The sample volume was then increased to 100µl containing 0.3M NaOAc pH 5.4, extracted once with 2FC and once with Chloroform before ethanol precipitation. Washed, dried samples were resuspended in 4µl of form stop buffer, heated to 90-100°C for 2 mins and loaded onto an acrylamide sequencing gel. End labelled fragments of DNA or a sequencing ladder were used as size markers.

First Strand Buffer	50mM Tris. Cl pH 8.3 at 42°C
	10mM $MgCl_2$
	140mM KCl
	10mM DTT
	1mM each dNTP
	0.75 units/µl RNasin

DNA Sequence Analysis.

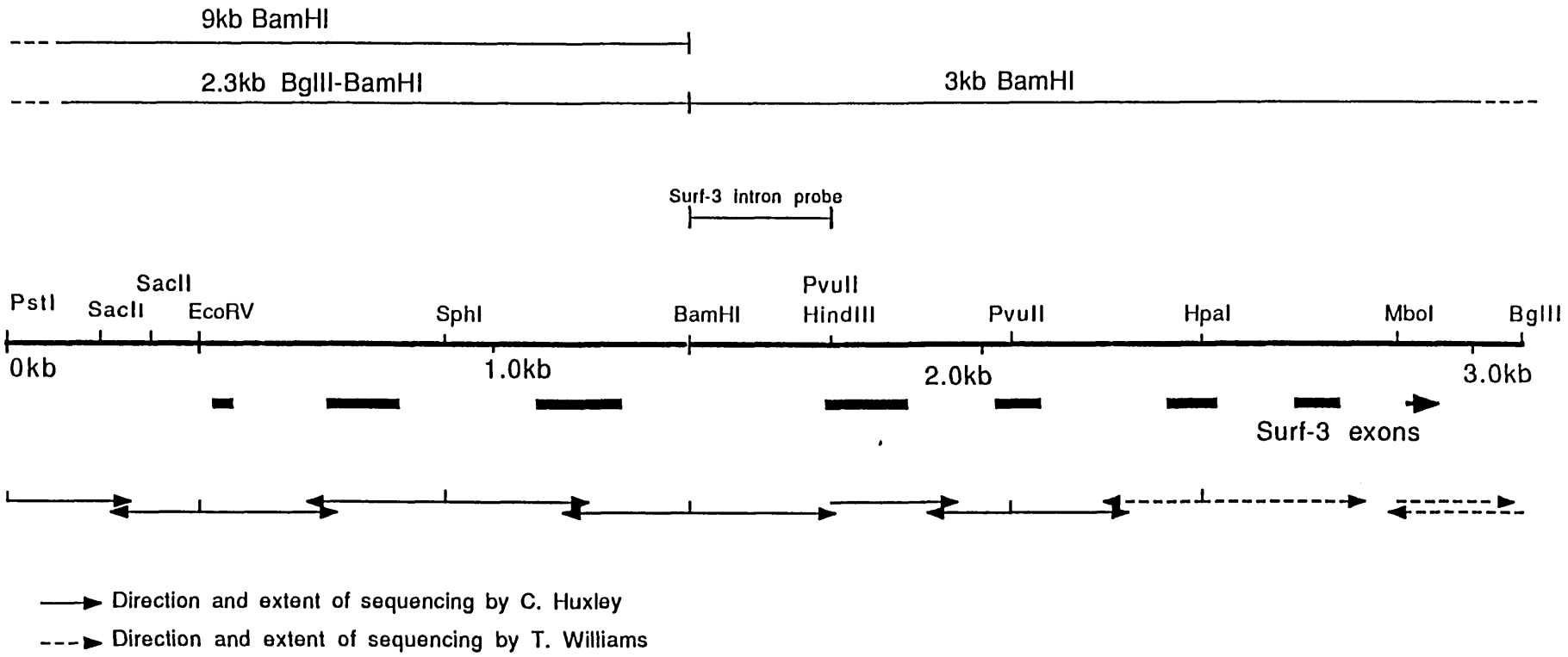
The majority of DNA sequence analysis was performed by the dideoxy chain termination method (Sanger et al 1977) using recombinant M13 templates grown in E.coli strain TG1 as detailed in Bankier and Barrell (1983), with the modification that templates were protein extracted twice with TE saturated phenol, and once with chloroform, before use. Sequenase kits were also used to sequence single stranded M13 DNA or double stranded bluescript

DNA, as described in the sequenase manual. End-labelled DNA was sequenced as described by Maxam and Gilbert (1980).

Computer Analysis of Sequence Data.

Primary sequence data were entered as VAX files using GENED (Intelligenetics) and were analysed using SEQ or PEP (Intelligenetics). The IFIND programme (Intelligenetics) was used to carry out Global nucleic acid homology searches against the EMBL (version 17.0) and GENBANK (version 58.0) databases. Protein homology searches were carried out using the IFIND programme (Intelligenetics) on the PIR (version 18.0) and Swiss Protein (version 9) databases.

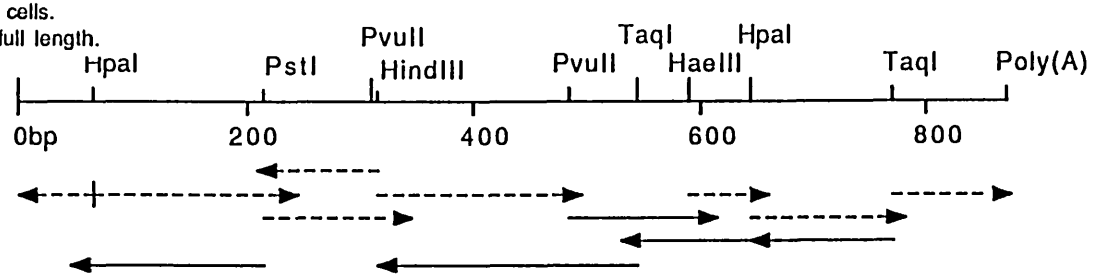
Appendix 1a: Map and Sequencing of Surf-3



Appendix 1b: Maps and Sequencing of Surf-3 cDNAs

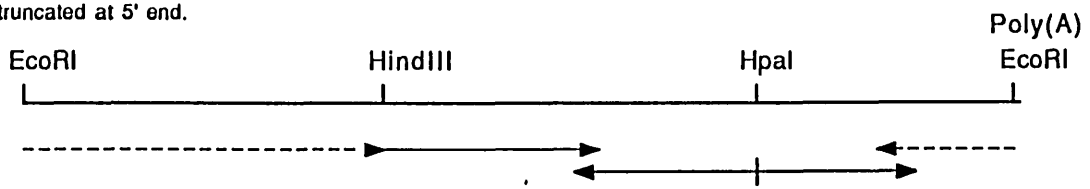
cDNA MY3

Cloned by T. Williams (1986)
Made from RNA from Balb/C A31 3T3 cells.
Cloned with EcoRI linkers into Pxf3, full length.



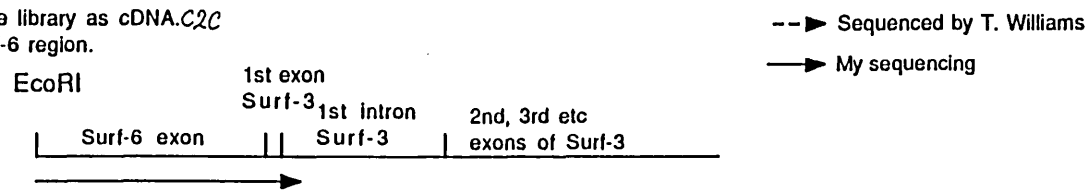
cDNA C2C

Cloned by T. Williams (1986)
Made from RNA from COS cells transfected with the19kb EcoRI fragment.
Cloned with EcoRI linkers into Pxf3, truncated at 5' end.

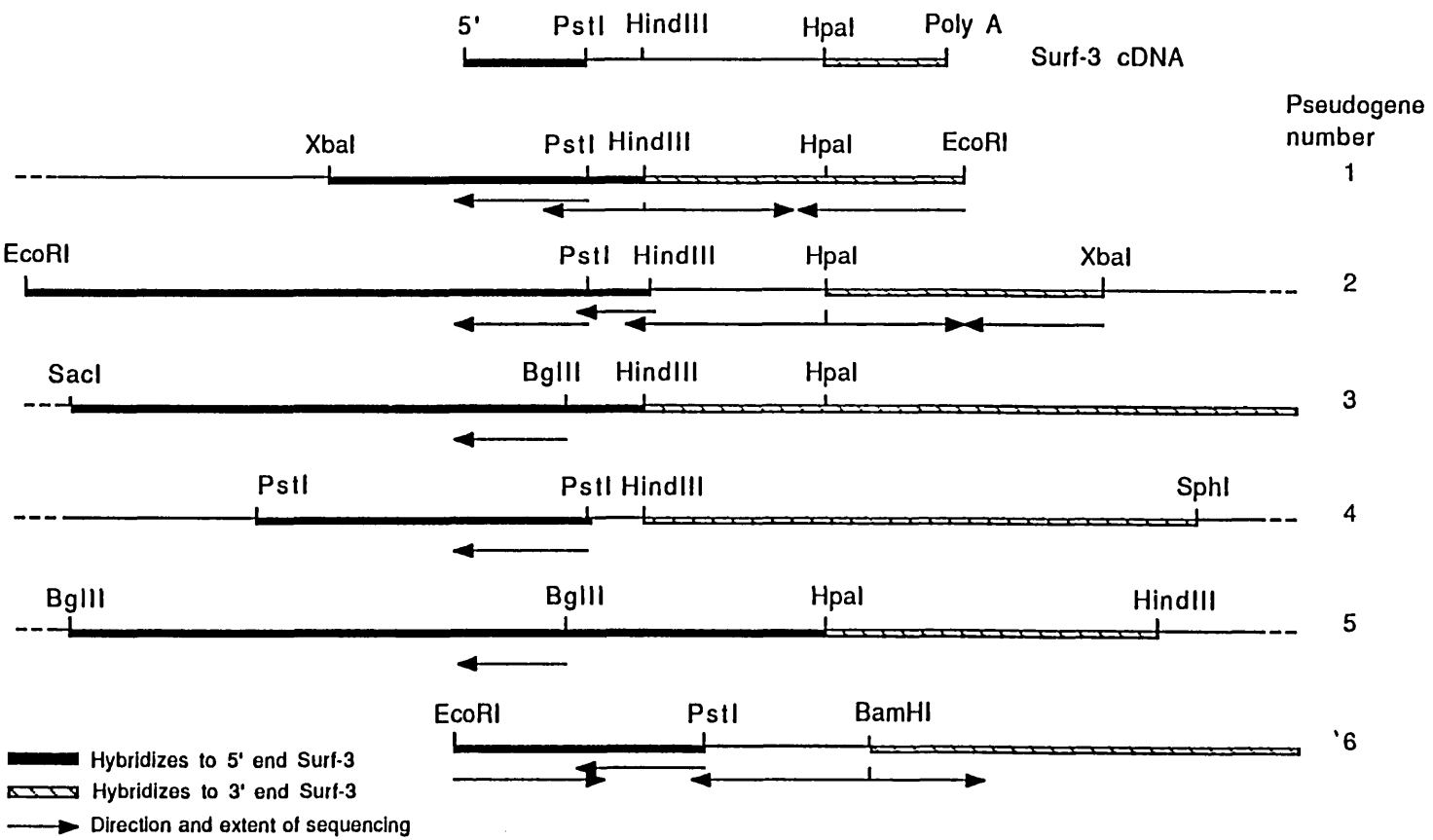


cDNA C3B

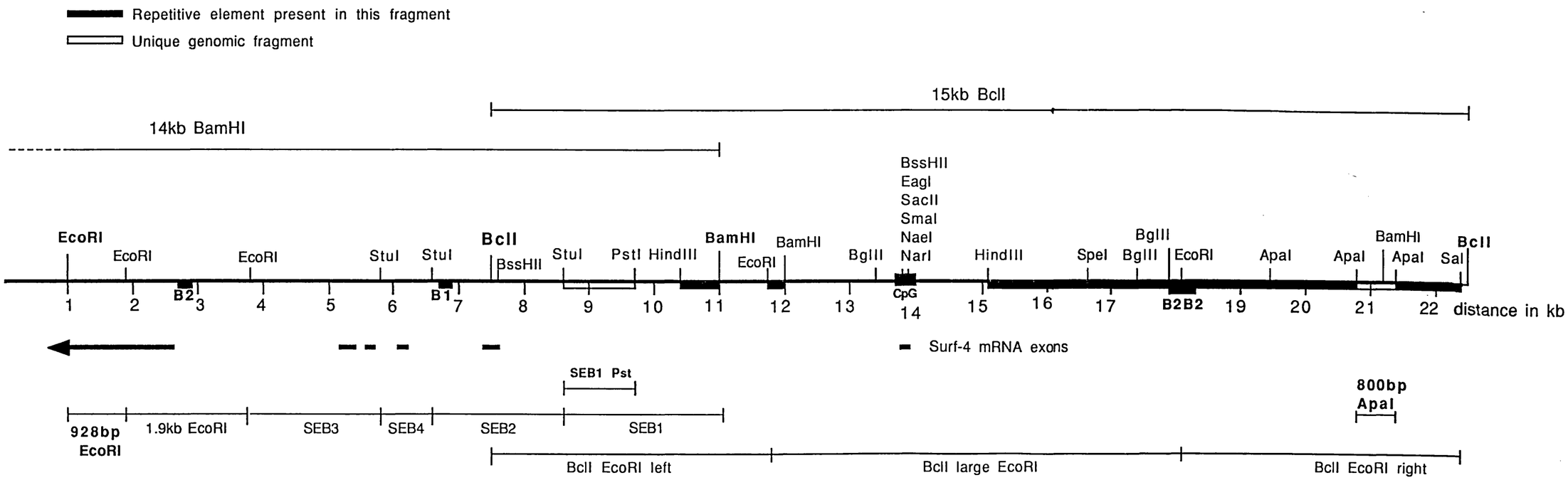
Cloned by T. Williams (1986), same library as cDNA *C2C*
The 5' end is derived from the Surf-6 region.



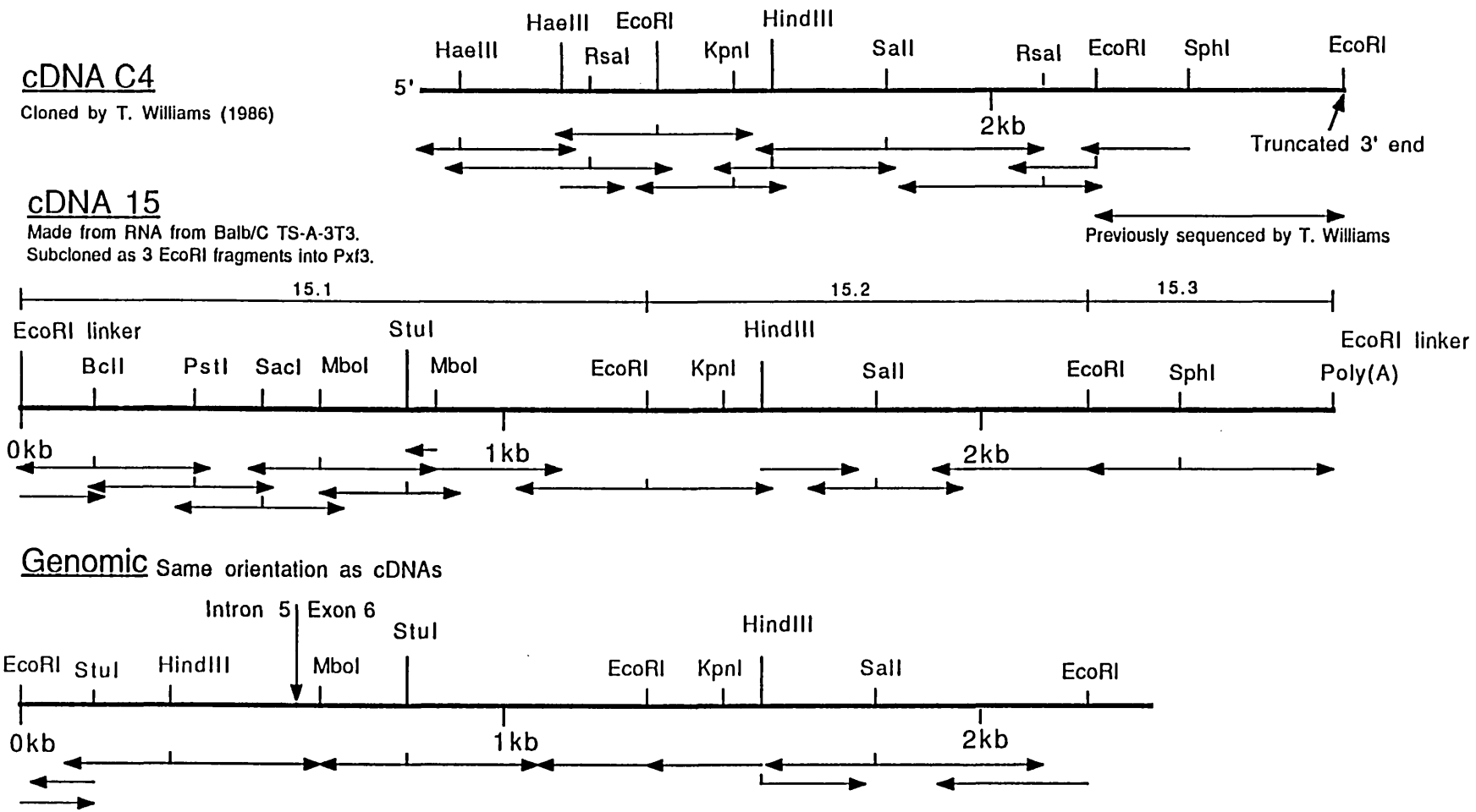
Appendix 1c: Maps and Sequencing of Surf-3 pseudogenes



Appendix 2a: Map of Surf-4



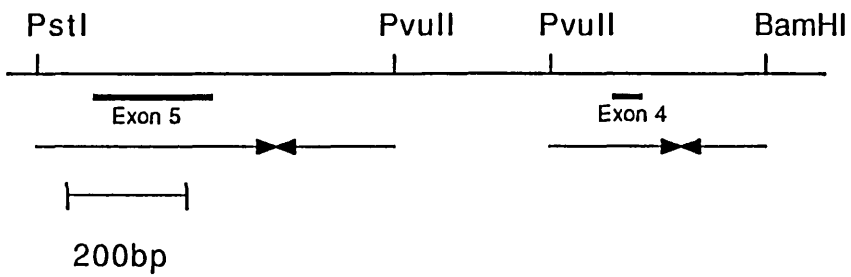
Appendix 2b: Map and Sequencing of Surf-4 cDNAs



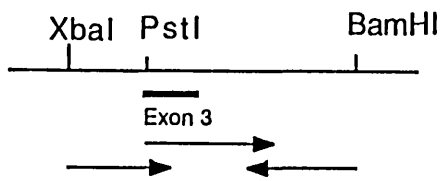
Appendix 2c: Maps and Sequencing of genomic exons of Surf-4

Same orientation as appendix 2a

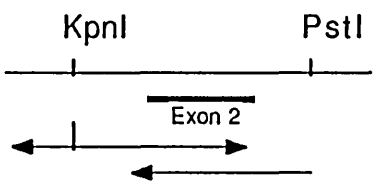
Exons 5 and 4, in fragment "SEB 3"



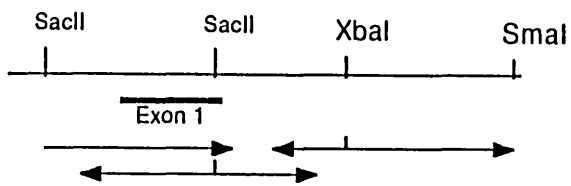
Exon 3, in fragment "SEB 4"



Exon 2, in fragment SEB 2



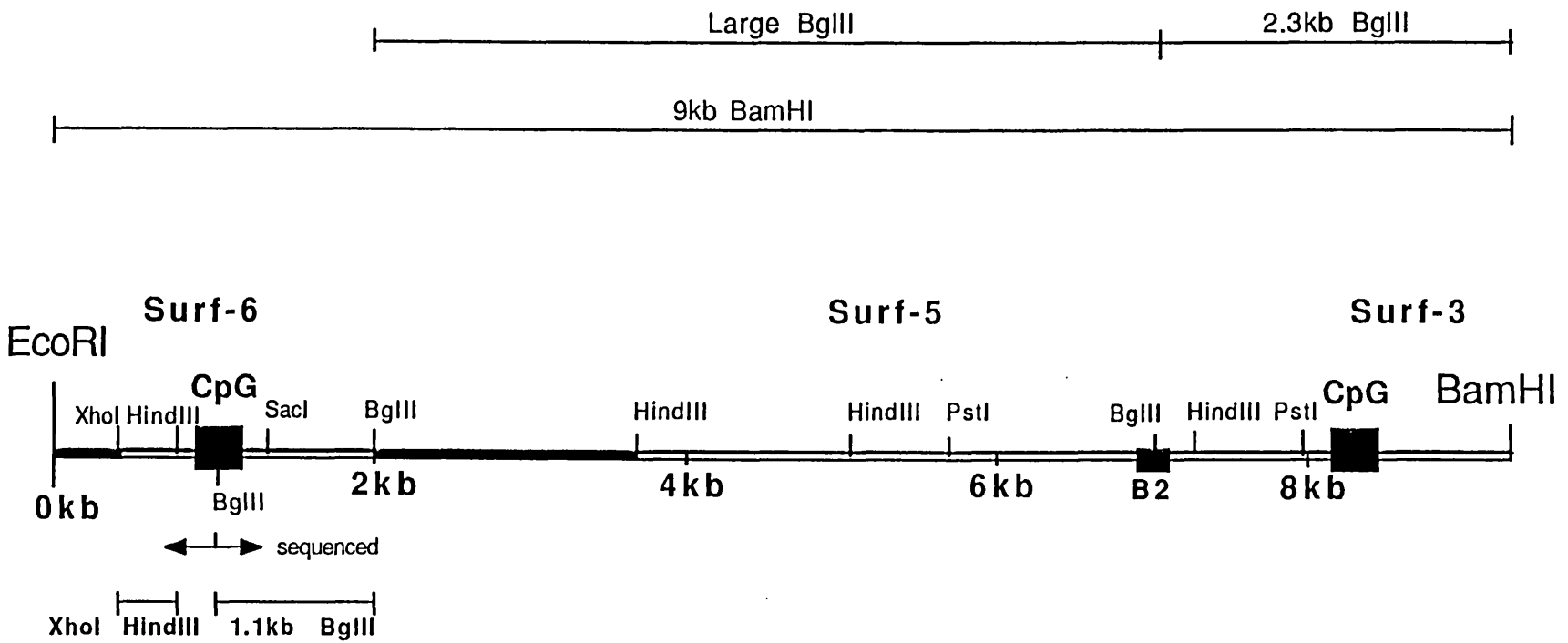
Exon 1, in fragment "BCII large EcoRI"



Appendix 3: Map of the Surf-5 and -6 region

indicates repetitive DNA

indicates unique DNA



Appendix 4a: Protein Consensus sequences

RNA binding site

An RNA binding consensus has been observed in a number of unrelated proteins (Preugschat 1988. Proc. Natl. Acad. Sci. USA 85: 9669-9673).

K/R-G-F/Y-G/A-F/Y-V/I-X-F/Y

(Swanson et al 1987 Mol. Cell Biol. 7: 1731-1739)

Zinc finger DNA binding domain

The zinc finger binding domain was first recognized in transcription factor IIIA of Xenopus laevis and has been identified in a wide range of DNA binding proteins (Berg 1986. Science. 232: 485-487)

$C/H\text{-}X_{2\text{-}5}\text{-}C/H\text{-}X_{2\text{-}15}\text{-}C/H\text{-}X_{2\text{-}5}\text{-}C/H$

Protein kinase A phosphorylation sites

cAMP dependent protein, activation enhances cellular differentiation and loss of proliferation (Trepel & Neckers 1987 Mol. Cell Biol. 7: 2644-2648)

R-R-X-S/T

(Edelman et al 1987 Ann. Rev. Biochem. 56: 567-613)

Casein Kinase II phosphorylation site

CKII activation is associated with serum-stimulated cell growth (Figge et al 1988 J. Virol. 62: 1814-1818)

S/T-X-X-E/D where X are often E/D and the sequence is proceeded and preceded by E/D (Kuenzel et al 1987 J. Biol. Chem. 262: 9136-9140)

affects SV40 LT, Adenovirus E1a protein, c-Myc.

Leucine zipper

Present in C/EBP and Myc, both of which are DNA binding proteins, this motif is thought to be important for dimerization.

$L\text{-}X_6\text{-}L\text{-}X_6\text{-}L$ etc, at least four leucines in an α-helix region, and interspersed with charged amino acids. (Landschulz et al 1988. Science 240: 1759-1764).

Also found in GCN4, Jun, Fos and CREB (Gonzalez et al 1989 Nature 337: 749-752)

Nucleotide binding site

A motif of nucleotide binding proteins (Walker et al 1982. EMBO 8: 945-951)

G/A-X-X-X-X-G-K-T/S

B motif of nucleotide binding proteins (Walker et al 1982. EMBO 8: 945-951)

R/K-X-X-X-G-X-X-X-L-HYDRO-HYDRO-HYDRO-HYDRO-D

Both the above form an ATP binding fold found in a variety of kinases.

The A motif is found in a wide variety of other ATP dependent proteins (Hodgman 1988 Nature 333: 22-23)

The A and B motifs are also found in a goup of eight eIF-4A like proteins (Linder et al 1989 Nature 337: 1211-122)

The B motif also takes the form

V/I-L-D-E-A-D-X-M/L-L-X-X-G-F

Another motif is also found in these proteins

H-R-I-G-R

Appendix 4b: Transcription factor binding sites

AP-1, Jun

C/GTGACTC/AA

47kd

hMT IIA, SV40 hCollagenase, TPA inducible

Lee et al 1987. Cell 49: 741-752

Vogt & Tjian. 1988. Oncogene 3: 3-7.

CTGACTCA

Angel et al 1987 Cell 49:729-739

AP-2

CCCCAGGC T/C C C/G C C A/C N G/C C/G G/C

52kd

SV40 enhancer, hMT IIA

Mitchell et al 1987 Cell 50:847-861

Imagawa et al 1987. Cell 51: 251-260

AP-3, EBP20

TGTGGA/TA/TA/TG

enhancer core, basal level activity

Johnson et al 1987. Genes & Devel. 1: 133-146.

Sassone-Corsi & Borelli. 1986. Trends. Genet. 2: 215-219

C-2

CATGTG

IgH enhancer

45kd

Peterson & Calame, Mol. Cell Biol. 1989. 9: 776-786

CCArGG box

CCA/TA/TA/TA/TA/TA/TGG

rat & chicken skeletal muscle actin genes, cytoplasmic β-actin genes, γ-actin gene

Minty & Kedes. 1986. Mol. Cell. Biol. 6: 2125-2136.

Miwa & Kedes. 1987. Mol. Cell. Biol. 7: 2803-2813.

CREB, E2AF, E3F1, E3F2, E4F, ETF-A, E4TF3, EIIaE-B, AP6, ATF, F2

TGACGTCA

cAMP response element

43kd cloned by Gonzalez et al 1989 Nature 337: 749-752

Montimy et al 1986 Proc Natl Acad Sci 83: 6682

Yamamoto et al 1988 Nature 334: 494-498

CTF, NF1, CP1, CP2, EBP

TGGC/AN$_5$GCCAA (CCAAT)

52-66kd

Jones et al 1987 Cell 48: 79

Chodosh et al 1988. Cell 53: 11-24.

E2 of BPV-1

ACCN6GGT

McBride et al 1988

E2F

TTTCGCGC

Reichel et al 1987 Cell 48: 501

HSTF

C--GAA--TTC--G

same factor in yeast and Drosophila

70kd

Wiederrecht et al 1987. Cell 48: 507-515.

MLTF, USF

GTAGGCCACGTGACCGG (CCACGTGA)

Sawadogo and Roeder 1985. Cell 43: 165-175.

NF-κB

GGGA/GA/CTT/CT/CCC

k immunoglobulin enhancer B site

Blanar et al Mol. Cell. Biol 1989. 9: 844-846.

Octamer, IgNF-A, NFIII, OTF, OBP, TRF

ATTTGCAT

Oct-2A	60kDa	lymphocyte specific
Oct-2B	75kDa	lymphocyte specific
Oct-1	100kDa	present in many cell types

all immunoglobulin heavy and light chain promoters.

Histone H2b gene promoter, cell cycle specific.

Fletcher et al 1988 Cell 773-781

also UsnRNA genes

Singh et al 1986

Scheidereit et al 1987 Cell 51: 783-793.

Ribosomal proteins, first exon

CG/TGCCATCT

R. Perry (personal communication)

RPG

AACAT/CCCG/ATG/ACATT/CT/A

Planta & Raue 1988 TIG 4: 64-68

Yeast ribosomal proteins

SP1

G/TG/AGGCGG/TG/AG/AC/T

105, 95kd, basal level transcription

Briggs et al 1986. Science 234: 47-52

Kadonaga,1986. TIBS 11: 20-23.

SRF, CArG

GATGTCCATATTAGGACATC (CCA/T_6GG)

67kd

Treisman 1986. Cell 46: 567-574.

Boxer et al 1989. Mol. Cell Biol. 9: 515-522.

TFIID

TATAT/AAT/A

TATA box, basal level transcription

Corden et 1980. Science 209: 1406-1414

TRE

AGGTAAG AGGGACG

Thyroid hormone response element, rat growth hormone promoter

Brent et al 1989. J. Biol. Chem. 264: 178-182

References

Aaronson, S. A. and G. J. Todaro. 1968. Development of 3T3-like Lines from Balb/c Mouse Embryo Cultures: Transformation Susceptibility to SV40. J. Cell. Physiol. **72:** 141-148.

Adelman, J. P., C. T. Bond, J. Douglass, E. Herbert. 1987. Two Mammalian Genes Transcribed from Opposite Strands of the Same DNA Locus. Science **235:** 1514-1517.

Adhya, S. and M. Gottesman. 1982. Promoter Occlusion: Transcription through a Promoter May Inhibit Its Activity. Cell **29:** 939-944.

Agrawal, M. G. and L. H. Bowman. 1987. Transcriptional and Translational Regulation of Ribosomal Protein Formation during Mouse Myoblast Differentiation. J. Biol. Chem. **262:** 4868-4875.

Arndt-Jovin, D. J. and T. M. Jovin. 1977. Analysis and Sorting of Living Cells According to Deoxyribonucleic Acid Content. J. Histochem. Cytochem. **25:** 585-589.

Atchison, M. L., O. Meyuhas and R. P. Perry. 1989. Localization of transcriptional regulatory elements and nuclear factor binding sites in the mouse ribosomal protein gene rpL32. Mol. Cell. Biol. submitted.

Avner, P., L. Amar, L. Dandolo and J. L. Guenet. 1988. Genetic analysis of the mouse using interspecific crosses. TIG **4:** 18-23.

Bankier, A. T. and B. G. Barrell. 1983. In Techniques in nucleic acid biochemistry. Flavell, R.A. (ed) Shotgun DNA sequencing. Elsevier Scientific, Limerick, Ireland pp 1-34.

Barlow, D. P., M. Bucan, H. Lehrach, B. L. M. Hogan and N. M. Gough. 1987. Close genetic and physical linkage between the murine haemopoietic growth factor genes GM-CSF and Multi-CSF (IL3). EMBO **6:** 617-623.

Barlow, D. P. and H. Lehrach. 1987. Genetics by gel electrophoresis: the impact of pulsed field gel electrophoresis on mammalian genetics. TIG **3:** 167-171.

Bass, B. L. and H. Weintraub. 1987. A Developmentally Regulated Activity That Unwinds RNA Duplexes. Cell **48:** 607-613.

Bass, B. L. and H. Weintraub. 1988. An Unwinding Activity That Covalently Modifies Its Double-Stranded RNA Substrate. Cell **55:** 1089-1098.

Beck, C. F. and R. A. J. Warren. 1988. Divergent Promoters, a Common Form of Gene Organization. Microb. Review **52:** 318-326

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival and F. Rodier. 1985. The Mosaic Genome of Warm-Blooded Vertebrates. Science **228:** 953-958.

Bernards, A., C. M. Rubin, C. A. Westbrook, M. Paskind and D. Baltimore. 1987. The First Intron in the Human c-abl Gene Is at Least 200 Kilobases Long and Is a Target for Translocations in Chronic Myelogenous Leukemia. Mol. Cell. Biol. **7:** 3231-3236.

Bernards, A. M. Paskind and D. Baltimore. 1988. Four murine *c-abl* mRNAs arise by usage of two transcriptional promoters and alternative splicing. Oncogene **2:** 297-304.

Biolabs Catalog. 1988. New England Biolabs, Inc. 32 Tozer Road, Beverly, MA. USA

Bird, A., M. Taggart, M. Frommer, O. J. Miller and D. Macleod. 1985. A Fraction of the Mouse Genome That Is Derived from Islands of Nonmethylated, CpG-Rich DNA. Cell **40:** 91-99.

Bird, A. P. 1986. CpG-rich islands and the function of DNA methylation. Nature **321:** 209-213.

Blattner, F. R., A. E. Blechl, K. Denniston-Thompson, H. E. Faber, J. E. Richards, J. L. Slightom, P. W. Tucker and O. Smithies. 1978. Cloning Human Fetal γ Globin and Mouse α-Type Globin DNA: Preparation and Screening of Shotgun Collections. Science **202:** 1279-1283.

Bonhomme, F., J. Catalan, J. Britton-Davidian, V. M. Chapman, K. Moriwaki, E. Nevo and L. Thaler. 1984. Biochemical Diversity and Evolution in the Genus *Mus*. Biochem. Genetics **22:** 275-303.

Bossy, B., L. M. C. Hall and P. Spierer. 1984. Genetic activity along 315 kb of the *Drosophila* chromosome. EMBO **3:** 2537-2541.

Bowman, L. H. 1987. The Synthesis of Ribosomal Proteins S16 and L32 is Not Autogenously Regulated during Mouse Myoblast Differentiation. Mol. Cell. Biol. **7:** 4464-4471.

Bozzoni, I., P. Fragapane, F. Annesi, P. Pierandrei-Amaldi, F. Amaldi and E. Beccari. 1984. Expression of Two *Xenopus laevis* Ribosomal Protein Genes in Injected Frog Oocytes. J. Mol. Biol. **180:** 987-1005.

Brown, W. R. A. and A. P. Bird. 1986. Long-range restriction site mapping of mammalian genomic DNA. Nature **322:** 477-481.

Brown, W. R. A. 1988. A physical map of the human pseudoautosomal region. EMBO **7:** 2377-2385.

Bucan, M., T. Yang-Feng, A. M. Colbderg-Poley, D. J. Wolgemuth, J. Guenet, U. Francke and H. Lehrach. 1986. Genetic and cytogenetic localisation of the homeo box containing genes on mouse chromosome 6 and human chromosome 7. EMBO **5:** 2899-2905.

Buratowski, S., S. Hahn, P. A. Sharp and L. Guarente. 1988. Function of a yeast TATA element-binding protein in a mammalian transcription system. Nature **334:** 37-42.

Burbelo, P. D., G. R. Martin and Y. Yamada. 1988. α1(IV) and α2(IV) collagen genes are regulated by a bidirectional promoter and a shared enhancer. Proc. Natl. Acad. Sci. USA **85:** 9679-9682.

Burmeister, M., A. P. Monaco, E. F. Gillard, G. B. ban Ommen, N. A. Affara, M. A. Ferguson-Smith, L. M. Kunkel and H. Lehrach. 1988. A 10-Megabase Physical Map of Human Xp21, Including the Duchenne Muscular Dystrophy Gene. Genomics **2:** 189-202.

Caput, D., B. Beutler, K. Hartog, R. Thayer, S. Brown-Shimer and A. Cerami. 1986. Identification of a common nucleotide sequence in the 3'-untranslated region of mRNA molecules specifying inflammatory mediators. Proc. Nat. Acad. Sci. USA **83:** 1670-1674.

Carroll, M. C., R. D. Campbell, D. R. Bentley and R. R. Porter. 1984. A molecular map of the human major histocompatibility complex class III region linking complement genes C4, C2 and factor B. Nature **307:** 237-240.

Casavant, N. C., S. C. Hardies, F. D. Funk, M. B. Comer, M. H. Edgell and C. A. Hutchison III. 1988. Extensive Movement of LINES ONE Sequences in β-Globin Loci of *Mus caroli* and *Mus domesticus*. Mol. Cell. Biol. **8:** 4669-4674.

Chaudhari, N and W. E. Hahn. 1983. Genetic Expression in the Developing Brain. Science **220:** 924-928.

Chen, C., T. Malone, S. K. Beckendorf and R. L. Davis. 1987. At least two genes reside within a large intron of the *dunce* gene of Drosophila. Nature **329:** 721-724.

Chen, I-T. and D. J. Roufa. 1988. The transcriptionally active human ribosomal protein S17 gene. Gene **70:** 107-116.

Chirgwin, J. M., A. E. Przybyla, R. J. MacDonald and W. J. Rutter. 1979. Isolation of Biologically Active Ribonucleic Acid from Sources Enriched in Ribonuclease. Biochem. **18:** 5294-5299.

Choi, O. B. and J. D. Engel. 1988. Developmental Regulation of β-Globin Gene switching. Cell **55:** 17-26.

Chung, S. and R. P. Perry. 1989. The importance of introns for expression of the mouse ribosomal protein gene rpL32. Mol. Cell Biol. submitted.

Church, G. M. and W. Gilbert. 1984. Genomic sequencing. Proc.Natl. Acad. Sci.USA **81:** 1991-1995.

Colberg-Poley, A., S. D. Voss, K. Chowdhury, C. L. Stewart, E. F. Wagner and P. Gruss. 1985. Clustered Homeo Boxes Are Differentially Expressed during Murine Development. Cell **43:** 39-45.

Corbin, V. and T. Maniatis. 1989. Role of transcriptional interference in the *Drosophila melanogaster* Adh promoter switch. Nature **337:** 279-282.

Cosman, D. 1987. Control of messenger RNA stability. Immunology Today **8:** 16-17.

Cullen, B. R., P. T. Lomedico and G. Ju. 1984. Transcriptional interference in avian retroviruses - implications for the promoter insertion model of leukaemogenesis. Nature **307:** 241-245.

Dadeva, M. D., M. A. Post-Beittenmiller and J. R. Warner. 1986. Proc. Natl. Adac. Sci. USA **83:** 5854-5857.

Dale, B. and B. Ozanne. 1981. Characterization of Mouse Cellular Deoxyribonucleic Acid Homologous to Abelson Murine Leukemia Virus-Specific Sequences. Mol. Cell. Biol. **1:** 731-742.

Daniels, G. R. and Deininger, P. L. 1985. Repeat sequence families derived from mammalian tRNA genes. Nature **317:** 819-822.

Doolittle, R. F. 1986. Of URFS and ORFS. University Science books, CA 94

Dudov, K. P. and R. P. Perry. 1984. The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron-containing genc and an unmutated processed gene. Cell **37:** 457-468.

Dudov, K. P. and R. P. Perry. 1986. Properties of a Mouse Ribosomal Protein Promoter. Proc. Natl. Acad. Sci. USA. **83:** 8545-8549.

Dynan, W. S. and R. Tjian. 1985. Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature **316:** 774-778.

Efrat, S. and D. Hanahan. 1987. Bidirectional Activity of the Rat Insulin II 5'-Flanking Region in Transgenic Mice. Mol. Cell. Biol. **7:** 192-198.

ElBaradi, T. T. A. L , C. A. F. M. van der Sande, W. H. Mager, H. A. Raue and R. J. Planta. 1986. The cellular level of yeast ribosomal protein L25 is controlled principally by rapid degradation of excess protein. Curr. Genet. **10:** 733-739.

Estivill, X., M. Farrall, P. J. Scambler, G. M. Bell, K. M. F. Hawley, N. J. Lench, G. P. Bates, H. C. Kruyer, P. A. Frederick, P. Stanier, E. K. Watson, R. Williamson and B. J. Wainwright. 1987. A candidate for the cystic fibrosis locus islolated by selection for methylation-free islands. Nature **326:** 840-845.

Featherstone, M. S., A. Baron, S. J. Gaunt, M. Mattei and D. Duboule. 1988. *Hox-5.1* defines a homeobox-containing gene locus on mouse chromosome 2. Proc. Natl. Acad. Sci. USA **85:** 4760-4764.

Fischel-Ghodsian, N., R. D. Nicholls and D. R. Higgs. 1987. Unusual features of CpG-rich (HTF) islands in the human α globin complex: association with non-functional pseudogenes and presence within the 3' portion of the ξ gene. Nucl. Acids. Res. 15: 9215-9225.

Fried, M., M. Griffiths, B. Davies, G. Bjursell, G. La Mantia, and L. Lania. 1983. Isolation of cellular DNA sequences that allow expression of adjacent genes. Proc. Natl. Acad. Sci. USA **80:** 2117-2121.

Frischauf, A., H. Lehrach, A. Pouska and N. Murray. 1983. Lambda Replacement Vectors Carrying Polylinker Sequences. J. Mol. Biol. **170:** 827-842

Gardiner-Garden, M. and M. Frommer. 1987. CpG-Islands in Vertebrate Genomes. J. Mol. Biol. **196:** 261-282.

Gasser, S. M. and U. K. Laemmli. 1987. A glimpse at chromosomal order. TIG **3:** 16-22.

Giallongo, A., J. Yon and M. Fried. 1989. Ribosomal Protein L7a Is Encoded by a Gene (*Surf-3*) within the Tightly Clustered Mouse Surfeit Locus. Mol. Cell. Biol. **9:** 224-231.

Gilbert, W. 1978. Why genes in pieces? Nature **271:** 501.

Glantz, S. A. 1987. Primer of biostatistics. McGraw-Hill Book Company. pp 101-137

Gluzman, Y. 1981. SV40-transformed simian cells support the replication of early SV40 mutants. Cell. **23:** 175-182.

Goldman, M. A., G. P. Holmquist, M. C. Gray, L. A. Caston and A. Nag. 1984. Replication Timing of Genes and Middle Repetitive Sequences. Science **224:** 686-692.

Grchnik, J. M., B. A. French and R. J. Schwartz. 1988. The Chicken Skeletal α-Actin Gene Promoter Region Exhibits Partial Dyad Symmetry and a Capacity To Drive Bidirectional Transcription. Mol. Cell. Biol. **8:** 4587-4597.

Grdina, D. J., M. L. Meistrich, R. E. Meyn, T. S. Johnson and R. A. White. 1987. Cell Synchrony Techniques, A Comparison of Methods. In Techniques in Cell Cycle Analysis: Gray, J. W. and Z. Darzynkiewicz (eds). Humana Press, Clifton, New Jersey. pp 367-402.

Greaves, D. R., F. D. Wilson, G. Lang and D. Kioussis. 1989. Human CD2 3'-Flanking Sequences Confer High-Level T Cell-Specific, Position-Independent Gene Expression in Transgenic Mice. Cell **56:** 979-986.

Green M. C. 1981. Genetic Variants and Strains of the Laboratory Mouse. Gustav Fischer Verlag, Stuttgart, NewYork. p 217

Green, P. J., O. Pines and M. Inouye. 1986. The Role of Antisense RNA in Gene Regulation. Ann Rev. Biochem. **55:** 569-597.

Grosveld, F., G. B. van Assendelft, D. R. Greaves, and G. Kollias. 1987. Position-Independent, High-Level Expression of the Human β-Globin Gene in Transgenic Mice. Cell **51:** 975-985.

Gubler, U. and B. J. Hoffman. 1983. A simple and very efficient method for generating cDNA libraries. Gene. **25:** 263-269.

Hanahan, D. 1983. Studies on Transformation of *Escherichia coli* with Plasmids. J. Mol. Biol. **166:** 557-580.

Hanahan, D. 1985. In DNA Cloning a Practical Approach **1**: Glover,D.M. (ed). IRL Press, Oxford, England. Techniques for Transformation of *E. coli.* 109-135.

Henikoff, S., M. A. Keene, K. Fechtel and J. W. Fristrom. 1986. Gene within a Gene: Nested Drosophila Genes Encode Unrelated Proteins on Opposite DNA Strands. Cell **44:** 33-42.

Herrmann, B., M. Bucan, P. E. Mains, A. Frischauf, L. M. Silver and H. Lehrach. Genetic Analysis of the Proximal Portion of the Mouse t Complex: Evidence for a Second Inversion within t Haplotypes. Cell **44:** 469-476.

Heywood, S. M. 1986. tcRNA as a naturally occurring antisense RNA in eukaryotes. Nucl. Acids Res. **14:** 6771-6772.

Hoffman, E. K., S. P. Trusko, N. Freeman and D. L. George. 1987. Structural and Functional Characterization of the Promoter Region of the Mouse c-Ki-*ras* Gene. Mol. Cell. Biol. **7:** 2592-2596.

Holmes, D. S. and M. Quigley. 1981. A Rapid Boiling Method for the Preparation of Bacterial Plasmids. Anal. Biochem. **114:** 193-197

Holmquist, G. P. 1987. Role of Replication Time in the Control of Tissue-specific Gene Expression. Am. J. Hum. Genet. **40:** 151-173.

Horowitz, H. and T. Platt. 1982. Regulation of transcription from tandem and convergent promoters. Nucl. Acids Res. **10:** 5447-5465.

Huynh, T. V., R. A. Young and R. W. Davis. 1985. In DNA Cloning a Practical Approach **1:** Glover,D.M. (ed). IRL Press, Oxford, England. Constructing and Screening cDNA Libraries in λgt10 and λgt11. pp 49-78

Huxley, C., T. Williams and M. Fried. 1988. One of the Tightly Clustered Genes of the Mouse Surfeit Locus is a Highly Expressed Member of a Multigene Family whose Other Members are Predominantly Processed Pseudogenes. Mol.Cell. Biol. **8:**3898-3905.

Izant, J. G. and H. Weintraub. 1985. Constitutive and Conditional Suppression of Exogenous and Endogeous Genes by Anti-Sense RNA. Science **229:** 345-352.

Jagadeeswaran, P., B. G. Forget and S. M. Weissman. 1981. Short Interspersed Repetitive DNA Elements in Eucaryotes: Transposable DNA Elements Generated by Reverse Transcription of RNA Pol III Transcripts?. Cell **26:** 141-142

Jenkins, N. A., N. G. Copeland, B. A. Taylor and B. K. Lee. Dilute (*d*) coat colour mutation of DBA/2J mice is associated with the site of integration of an ecotropic MuLV genome. Nature **293:** 370-374.

John, B. and G. Miklos. 1988. The Eukaryote Genome in Development and Evolution. Allen & Unwin, London.

Johnston, L, A., M. A. Kotarski, D. J. Jerry and L. P. Kozak. 1989. An Ubiquitously Expressed Gene 3.5 Kilobases Upstream of the Glycerol-3-Phosphate Dehydrogenase Gene in Mice. Mol. Cell. Biol **9:** 935-945.

Kadonaga, J. T., K. A. Jones and R. Tjian. 1986. Promoter-specific activation of RNA polymerase II transcription by Sp1. TIBS **11:** 20-23

Katzir, N., A. Oppenhein, M. Belfort and A. B. Oppenheim. 1976. Activation of the Lambda *int* Gene by the cII and cIII Gene Products. Virology **74:** 324-331.

Kalb, V. F., S. Glasser, D. King and J. B. Lingrel. 1983. A cluster of repetitive elements within a 700 base pair region in the mouse genome. Nucl. Acids Res. **11:** 2177-2184.

Kidd, S. and M. W. Young. 1986. Transposon-dependent mutant phenotypes at the *Notch* locus of Drosophila. Nature **323:** 89-91.

Kim, S. K. and B. J. Wold. 1985. Stable Reduction of Thymidine Kinase Activity in Cells Expressing High Levels of Anti-Sense RNA. Cell **42:** 129-138.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press pp. 181-183.

Koenig, M., E. P. Hoffman, C. J. Bertelson, A. P. Monaco, C. Feener and L. M. Kunkel. 1987. Complete Cloning of the Duchenne Muscular Dystrophy (DMD) cDNA and Preliminary Genomic Organization of the DMD Gene in Normal and Affected Individuals. Cell **50:** 509-517.

Koenig, R. J., M. A. Lazar, R. A. Hodin, G. A. Brent, P. R. Larsen, W. W. Chin and D. D Moore. 1989. Inhibition of thyroid hormone action by a non-hormone binding c-erbA protein generated by alternative mRNA splicing. Nature **337:** 659-661.

Kollias, G., N. Wrighton, J. Hurst and F. Grosveld. 1986. Regulated Expression of Human Aγ-, β-, and Hybrid γβ-Globin Genes in Transgenic Mice: Manipulation of the Developmental Expression Patterns. Cell **46:** 89-94.

Kollias, G., J. Hurst, E. deBoer and F. Grosvelt. 1987. The human β-globin gene contains a downstream developmental specific enhancer. Nucl. Acids Res. **15:** 5739-5747.

Kozak, M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs Nucl.Acid.Res **15:** 8125-8148.

Kozma, S. C., S. M. S. Redmond, F. Xiao-Chang, S. M. Saurer, B. Groner and N. E. Hynes. 1988. Activation of the receptor kinase domain of the trk oncogene by recombination with two different cellular sequences. EMBO. **7:** 147-154

Krieg, P. A. and D. A. Melton. 1984. Formation of the 3' end of histone mRNA by post-transcriptional processing. Nature **308:** 203-206.

Kruys, V. I., M. G. Wathelet and G. A. Huez. 1988. Identification of a translation inhibitory element (TIE) in the 3' untranslated region of the human interferon-β mRNA. Gene **72:** 191-200.

Lalley, P. A., M. T. Davisson, J. A. M. Graves, S. J. O'Brien, T. H. Roderick, D. P. Doolittle and A. Hillyard. 1988. Report of the committee on comparative mapping. Human Gene Mapping 9.5: Update to the Ninth International Workshop on Human Gene Mapping. Cytogenet Cell Genet **49:** 227-235

Lavia, P., D. Macleod and A. Bird. 1987. Coincident start sites for divergent transcripts at a randomly selected CpG-rich island of mouse. EMBO **6:** 2773-2779.

Lazar, M. A., R. A. Hodin, D. S. Darling and W. W. Chin. 1989. A Novel Member of the Thyroid/Steroid Hormone Receptor Family Is Encoded by the Opposite Strand of the Rat c-*erb*Aα Transcriptional Unit. Mol. Cell. Biol. **9:** 1128-1136.

Leff, S. E., M. G. Rosenfeld. 1986. Complex Transcriptional Units: Diversity in Gene Expression by Alternative RNA Processing. Ann. Rev. Biochem. **55:** 1091-1117.

Lindsay, S. and A. P. Bird. 1987. Use of restriction enzymes to detect potential gene sequences in mammalian DNA. Nature **327:** 336-338.

Lyon, M. F. 1988a. The William Allan Memmorial Award Address: X-Chromosome Inactivation and the Location and Expression of X-linked Genes. Am. J. Hum. Genet. **42:** 008-016.

Lyon, M. F. 1988b. Mouse Chromosome Atlas. Mouse News Letter **81**

Maeda, N. and O. Smithies. 1986. The evolution of multigene families: Human Haptoglobin Genes. Ann. Rev. Genet. **20:** 81-108.

Mager, W. H. 1988. Control of ribosomal protein gene expression. Biochemica et Biophysica Acta. **949:** 1-15.

Maniatis, T., E. F. Fritsch, J. Lauer and R. M. Lawn. 1980. The Molecular Gentics of Human Hemoglobins. Ann. Rev. Genet. **14:** 145-178.

Maniatis, R., E. F. Fritsch and J. Sambrook. 1982. Molecular cloning, a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Martin-Zanca, D., S. H. Hughes and M. Barbacid. 1986. A human oncogene formed by the fusion of truncated tropomyosin and protein tyrosine kinase sequences. Nature. **319:** 743-748.

Marzluff, W. F., R. A. Graves. 1984. In Histone Genes: G. S. Stein, J. L. Stein and W. F. Marzluff (eds.) John Wiley & Sons.

Maxam, A. M. and M. Gilbert. 1980. Methods in Enzymology **65:** 499-560.

Meistrich, M. L., R. E. Meyn and B. Barlogie. 1977. Synchronization of Mouse L-P59 Cells by Centrifugal Elutriation Separation. Exp. Cell Res. **105:** 169-177.

Mellor, A. L. 1986. In Oxford Surveys on Eukaryotic Genes: N. Maclean (ed) Oxford University Press. p 102.

Meyn, R. E., R. R. Hewitt and R. M. Humphrey. 1975. Evaluation of S Phase Synchronization by Analysis of DNA Replication in 5-Bromodeoxyuridine. In Methods in Cell Biology: D. M. Prescott (ed) Academic, San Francisco. Vol IX pp103-113.

Meyuhas, O., E. A. Thompson, Jr. and R. P. Perry. 1987. Glucocorticoids Selectively Inhibit Translation of Ribosomal Protein mRNas in P1798 Lymphosarcoma Cells. Mol. Cell. Biol. **7:** 2691-2699.

Milner, R. J., C. Lai, K. A. Nave. D. Lenoir, J. Ogata and J. G. Sutcliffe. 1985. Nucleotide sequences of two mRNAs for rat brain myelin proteolipid protein. Cell. **42:** 931-939.

Mitchell, P. J., A. M. Carothers, J. H. Han, J. D. Harding, E. Kas, L. venolia and L. A. Chasin. 1986. Multiple Transcription Start Sites, DNase I-Hypersensitive Sites, and an Opposite-Strand Exon in the 5' Region of the CHO *dhfr* Gene. Mol. Cell. Biol. **6:** 425-440.

Mitchell, P. J., C. Wang, and R. Tjian. 1987. Positive and Negative Regulation of Transcription In Vitro: Enhancer-Binding Protein AP-2 Is Inhibited by SV40 T Antigen. Cell **50:** 847-861.

Mizuno, T., M. Chou and M. Inouye. 1984. A unique mechanism regulating gene expression: Translational inhibition by a complementary RNA transcript (micRNA). Proc. Natl. Acad. Sci. USA **81:** 1966-1970.

Moore, G. E., A. A. Sandberg and K. Ulrich. 1966. Suspension Cell Culture and *In Vivo* and *In Vitro* Chromosome Constitution of Mouse Leukemia L1210. J. Nat. Cancer Inst. **36:** 405-413.

Mount, S. M. 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10:** 459-472.

Nadeau, J. H. and B. A. Taylor. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. USA **81:** 814-818.

Nelson, M and M. McClelland. 1987. The effect of site-specific methylation on restriction-modification enzymes. Nucl. Acids Res. **15:** r219-r230.

Nickol, J. M. and G. Felsenfeld. 1988. Bidirectional control of the chicken β- and ε- globin genes by a shared enhancer. Proc. Natl. Acad. Sci. USA **85:** 2548-2552.

Nusse, R. 1988. The int genes in mammary tumorigenesis and in normal development. TIG **4:** 291-295.

Ohno, S. 1985. Dispensable genes. TIG **1:** 160-164.

Okamoto, K. and M. Freundlich. 1986. Mechanism for the autogenous control of the *crp* operon: Transcriptional inhibition by a divergent RNA transcript. Proc. Natl. Acad. Sci. USA **83:** 5000-5004.

Okano, H., K. Ikenaka and K. Mikoshiba. 1988. Recombination within the upstream gene of duplicated myelin basic protein genes of *myelin deficient shi*^{m1d} mouse results in the production of antisense RNA. EMBO **7:** 3407-3412.

Otsuka, A. and J. Abelson. 1978. The regulatory region of the biotin operon in *Escherichia coli.* Nature **276:** 689-694.

Page, D. C., R. Mosher, E. M. Simpson, E. M. C. Fisher, G. Mardon, J. Pollack, B. McGillivray, A. de la Chapelle and L. G. Brown. 1987. The Sex-Determining Region of the Human Y Chromosome Encodes a Finger Protein. Cell **51:** 1091-1104.

Palmiter, R. D. and R. L. Brinster. 1986. Germ-line Transformation of Mice. Ann. Rev. Genet. **20:** 465-499.

Petit, C., J. Levilliers and J. Weissenbach. 1988. Physical mapping of the human pseudo-autosomal region: comparison with genetic linkage map. EMBO **7:** 2369-2376.

Planta, R. J. and H. A. Raue. 1988. Control of ribosome biogenesis in yeast. TIG **4:** 64-68.

Pohl, T. M., M. Zimmer, M. E. MaDonald, B. Smith, M. Bucan, A. Poustka, S. Volinia, S. Searle, G. Zehetner, J. J. Wasmuth, J. Gusella, H. Lehrach and A. -M. Frischauf. 1988. Construction of a NotI linking library and isolation of new markers close to the Huntington's disease gene. Nucl. Acids. Res. **16:** 9185-9192.

Polisky, B. 1988. ColE1 Replication Control Circuitry: Sense from Antisense. Cell **55:** 929-932.

Poschl, E., R. Pollner and K. Kuhn. 1988. The genes for the α1(IV) and α2(IV) chains of human basement membrane collagen type IV are arranged head-to-head and separated by a bidirectional promoter of unique structure. EMBO **7:** 2687-2695.

Pritchard, C. A., P. J. Goodfellow and P. N. Goodfellow. 1987. Mapping the limits of the human pseudoautosomal region and a candidate sequence for the male-determining gene. Nature **328:** 273-275

Proudfoot, N. J. 1986. Transcriptional interference and termination between duplicated α-globin gene constructs suggests a novel mechanism for gene regulation. Nature **322:** 562-565.

Pruss, G. J. and K. Drlica. 1989. DNA Supercoiling and Prokaryotic Transcription. Cell **56:** 521-523.

Ptashne, M. 1986. A genetic switch. Blackwell Scientific Publications, Palo Alto, Calif.

Rappold, G. A., L. Stubbs, S. Labeit, R. B. Crkvenjakov and H. Lehrach. 1987. Identification of a testis-specific gene from the mouse t-complex next to a CpG-rich island. EMBO **6:** 1975-1980.

Rappold, G. A. and H. Lehrach. 1988. A long range restriction map of the pseudoautosomal region by partial digest PFGE analysis from the telomere. Nucl.Acids Res. **16:** 5361-5377.

Rebagliati, M. R. and D. A. Melton. 1987. Antisense RNA Injections in Fertilized Frog Eggs Reveal an RNA Duplex Unwinding Activity. Cell **48:** 599-605.

Reeves, R., T. S. Elton, M. S. Nissen, D. Lehn and K. R. Johnson. 1987. Posttranscriptional gene regulation and specific binding of the nonhistone protein HMG-I by the 3' untranslated region of bovine interleukin 2 cDNA. Proc. Natl. Acad. Sci. USA **84:** 6531-6535.

Rhoads, D.D., A. Dixit and D. J. Roufa. 1986. Primary Structure of Human Ribsomal Protein S14 and the Gene that Encodes it. **6:** 2774-2783.

Robert, B., P. Barton, A. Minty, P. Daubas, A. Weydert, F. Bonhomme, J. Catalan, D. Chazottes, J. Guenet and M. Buckingham. 1985. Investigation of genetic linkage between myosin and actin genes using an interspecific mouse back-cross. Nature. **314:** 181-183.

Roberts, W. M., A. T. Look, M. F. Roussel and C. J. Sherr. 1988. Tandem Linkage of Human CSF-1 Receptor (*c-fms*) and PDGF Receptor Genes. Cell **55:** 655-661.

Roebroek, A. J. M., J. A. Schalken, J. A. M. Leunissen, C. Onnekink, H. P. J. Bloemers and W. J. M. Van de Ven. 1986. Evolutionary conserved close linkage of the *c-fes/fps* proto-oncogene and genetic sequences encoding a receptor-like protein. EMBO **5:** 2197-2202.

Rogers, D. T., J. M. Lemire and K. A. Bostian. 1982. Acid phosphatase polypeptides in *Saccharomyces cerevisiae* are encoded by a differentially regulated multigene family. Proc. Natl. Acad. Sci. USA. **79:** 2157-2161.

Rogers, J. H. 1985. The Origin and Evolution of Retroposons. Int. Rev. Cyt. **93:** 187-279.

Rotenberg, M. O., M. Moritz and J. L. Woolford, Jr. 1988. Depletion of *Saccharomyces cerevisiae* ribosomal protein L16 causes a decrease in 60S ribosomal subunits and formation of half-mer polyribosomes. Genes & Development **2:** 160-172.

Ruley, H. E., L. Lania, F. Chaudry and M. Fried. 1982. Use of a cellular polyadenylation signal by viral transcripts in polymoma virus transformed cells. Nucleic Acids Res. **10:** 4515-4524.

Russell, G. J., P. M. B. Walker, R. A. Elton and J. H. Subak-Sharpe. 1976. Doublet Frequency Analysis of Fractionated Vertebrate Nuclear DNA. J. Mol. Biol. **108:** 1-23.

Saffer, J. D.and M. F. Singer. 1984. Transcription from SV40-like monkey DNA sequences. Nucl. Acids Res. **12:** 4769-4789.

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

Sassone-Corsi, P. and E. Borrelli. 1986. Transcriptional regulation by *trans*-acting factors. Trends Genet. **2:** 215-219.

Sawadogo, M. and R. G. Roeder. 1985. Interaction of a Gene-Specific Transcription Factor with the Adenovirus Major Late Promoter Upstream of the TATA Box Region. Cell **43:** 165-175.

Schulz, R. A. and B. A. Butler. 1989. Overlapping genes of *Drosophila melanogaster*: organization of the *z600-gonadal-Eip28/29* gene cluster. Genes & Development **3:** 232-242.

Schumperli, D. 1988. Multilevel regulation of replication-dependent histone genes. TIG **4:** 187-191.

Serfling, E., M. Jasin and W. Schaffner. 1985. Enhancers and eukaryotic gene transcription TIG **1:** 224-230.

Shaw, G. and R. Kamen. 1986. A Conserved AU Sequence from the 3' Untranslated Region of GM-CSF mRNA Mediates selective mRNA Degradation. Cell **46:** 659-667.

Simons, R. W. and N. Kleckner. 1983. Translational Control of IS10 Transposition. Cell **34:** 683-691.

Skowronski, J. and M. F. Singer. 1986. The Abundant LINE-1 Family of Repeated DNA Sequences in Mammals: Genes and Pseudogenes. Cold Spring Harbor Symposia on Quantitative Biology **LI:** 457-464.

Spencer, C. A., R. D. Gietz and R. B. Hodgetts. 1986. Overlapping transcription units in the dopa decarboxylase region of *Drosophila*. Nature **322**: 279-281.

Stein, G. S., F. Sierra, M. Plumb, F. Marashi, L. Baumbach, J. L. Stein, N. Carozzi and K. Prokopp. 1984. Organization and Expression of Human Histone Genes. In Histone Genes: Stein, G. S., J. L. Stein and W. F. Marzluff (eds). John Wiley & Sons, Inc pp 397-432.

Stevenson, B. J., O. Hagenbuchle and P. K. Wellauer. 1986. Sequence organization and transcriptional regulation of the mouse elastase II and trypsin genes. Nucl. Acids Res. **14**: 8307-8330.

Stylianos, E. A. and Kazazian, H. H. Jr. 1988. The molecular basis of hemophilia A in man. TIG **4**: 233-237.

Takahashi, M., M. C. Yoshida, H. Satoh, J. Hilgers, Y. Yaoita and T. Honjo. 1989. Chromosomal Mapping of the Mouse IL-4 and Human IL-5 Genes. Genomics **4**: 47-52.

Talbot, D., P. Collis, M. Antoniou, M. Vidal, F. Grosveld and D. R. Greaves. 1989. A dominant control region from the human β-globin locus conferring integration site-independent gene expression. Nature **338**: 352-355.

Taylor, B. A. 1978. In Origins of Inbred Mice: Morse, H. C. III (ed). Academic Press, NY, pp. 423-438.

Terasima, T. and L. J. Tolmach. 1961. Changes in X-ray Sensitivity of HeLa Cells during the Division Cycle. Nature **190**: 1210-1211.

Thomas, M. S. and M. Nomura. 1987. Translational regulation of the L11 ribosomal protein operon of *Escherichia coli*: mutations that define the target site for repression by L1. Nucl. Acids Res. **15**: 3085-3097.

Threadgill, D. W. and J. E. Womack. 1988. Regional Localization of Mouse *Abl* and *Mos* Proto-oncogenes by *in Situ* Hybridization. Genomics **3**: 82-86.

Tooze, J. 1981. DNA Tumour Viruses. In Molecular Biology of Tumour Viruses, part 1. Cold Spring Harbor Laboratory, New York, USA.

Travers, A. 1984. Conserved features of coordinately regulated *E. coli* promoters. Nucl. Acids Res. **12**: 2605-2618.

Tushinski, R. J. and J. R. Warner. 1982. Ribosomal Proteins Are Synthesized Preferentially in Cells Commencing Growth. J. Cell. Phsiol. **112**: 128-135.

Ullu, E. and C. Tschudi. 1984. *Alu* sequences are processed 7SL RNA genes. Nature **312**: 171-172

van Assendelft, G. B., O. Hanscombe, F. Grosveld and D. R. Greaves. 1989. The β-Globin Dominant Control Region Activates Homologous and Heterologous Promoters in a Tissue-Specific Manner. Cell **56**: 969-977.

van der Krol, A. R., J. N. M. Mol and A. R. Stuitje. 1988. Modulation of Eukaryotic Gene Expression by Complementary RNA or DNA Sequences. BioTechniques **6**: 958-976.

Varmus, H. E. 1984. The Molecular Genetics of Cellular Oncogenes. Ann. Rev. Genet. **18:** 553-612.

Wagner, M. and R. P. Perry. 1985. Characterization of the Multigene Family Encoding the Mouse S16 Ribosomal Protein: Stategy for Distinguishing an Expressed Gene from its Processed Pseudogene Counterparts by an Analysis of Total Genomic DNA. Mol. Cell. Biol. **5:** 3560-3576.

Wagner, R. W. and K. Nishikura. 1988. Expression of an RNA Duplex Unwindase Activity in Mammalian Cells. In Antisense RNA and DNA. Melton D. A. ed. Cold Spring Harbor Laboratory Current Communication. p 103

Ward, D. F. and N. E. Murray. 1979. Convergent Transcription in Bacteriophage λ: Interference with Gene Expression. J. Mol. Biol. **133:** 249-266.

Waye, J. S., S. J. Durfy, D. Pinkel, S. Kenwrick, M. Patterson, K. E. Davies and H. F. Willard. 1987. Chromosome-Specific Alpha Satellite DNA from Human Chromosome 1: Hierarchical Structure and Genomic Organization of a Polymorphic Domain Spanning Several Hundred Kilobase pairs of Centromeric DNA. Genomics **1:** 43-51.

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

Weiner, A. M., P. L. Deininger and A. Efstratiadis. 1986. Nonviral Retroposons: Genes, Pseudogenes, and Transposable Elements Generated by the Reverse Flow of Genetic Information. Ann. Rev. Biochem. **55:** 631-661.

West, R. W. Jr., R. R. Yocum and M. Ptashne. 1984. *Saccharomyces cerevisiae GAL1-GAL10* Divergent Promoter Region: Location and Function of the Upstream Activating Sequence UAS_G. Mol. Cell. Biol. **4:** 2467-2478.

Wiedemann, L. M. and R. P. Perry. 1984. Characterization of the Expressed Gene and Several Processed Pseudogenes for the Mouse Ribosomal Protein L30 Gene Family. Mol. Cell. Biol. **4:** 2518-2528.

Wiedemann, L. M., P. D'Eustachio, D. E. Kelley and R. P. Perry. 1987. Three Functional Ribosomal Protein Genes Are Unlinked in Mouse Genome. Somat cell genet. **13:** 77-80.

Williams, T. 1986. The Analysis of a Transcriptionally Active Region Surrounding a Mouse Enhancer Sequence. Thesis.

Williams, T. and M. Fried. 1986a. The MES-1 murine enhancer element is closely associated with the heterogeneous 5' ends of two divergent transcription units. Mol. Cell. Biol. **6:** 4558-4569.

Williams, T. and M. Fried. 1986b. A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends. Nature. **322:** 275-279.

Williams, T., J. Yon, C. Huxley and M. Fried. 1988. The mouse surfeit locus contains a very tight cluster of four "housekeeping" genes which is conserved through evolution. Proc. Natl. Acad. Sci. USA. **85:** 3527-3530.

Wilson, T. and R. Treisman. 1988. Removal of poly(A) and consequent degradation of *c-fos* mRNA facilitated by 3' AU-rich sequences. Nature **336**: 396-399.

Wilson, V. L. and P. A. Jones. 1983. DNA Methylation Decreases in Aging but Not in Immortal Cells. Science **220**: 1055-1057.

Woodcock, D. M., P. J. Crowther, S. Jefferson and W. P. Diver. 1988. Methylation at dinucleotides other than CpG: implications for human maintenance methylation. Gene **74**: 151-152.

Woudt, L. P., A. B. Smit, W. H. Mager and R. J. Planta. 1986. Conserved sequence elements upstream of the gene encoding yeast ribosomal protein L25 are involved in transcription activation. EMBO **5**: 1037-1040.

Wu, L., B. J. Morley and R. D. Campbell. 1987. Cell-Specific Expression of the Human Complement Protein Factor B Gene: Evidence for the Role of Two Distinct 5'-Flanking Elements. Cell **48**: 331-342.

Yokoyama, K. and F. Imamoto. 1987. Transcriptional control of the endogenous MYC protooncogene by antisense RNA. Proc. Natl. Acad. Sci. USA **84**: 7363-7367.

Yon, J., R. W. Palmer, D. Sheer and M. Fried. 1989. Localization of the Surfeit gene cluster containing the ribosomal protein gene L7a to chromosome bands 9q33-34. Ann. Hum. Genet. **53**: in press

Young, P. R., R. W. Scott, D. H. Hamer and S. J. Tilghman. 1982. Construction and expression in vivo of an internally deleted mouse α-fetoprotein gene: presence of a transcribed Alu-like repeat within the first intervening sequence. Nucl. Acids. Res. **10**: 3099-3117.