Expression and Functional Analysis of the Transcription Factor DMAHP

by

Sarah Elizabeth Harris

Thesis submitted for the degree of Doctor of Philosophy
University of Glasgow

Division of Molecular Genetics
Institute of Biomedical and Life Sciences

February 1999

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Myotonic dystrophy (DM) is the most prevalent form of adult muscular dystrophy and is characterised by myotonia and muscle weakness and atrophy. The genetic defect associated with DM is an unstable $(CTG)_n$ repeat in the 3' untranslated region (UTR) of a protein kinase gene *myotonic dystrophy protein kinase* (*DMPK*) and the promoter region of a homeobox gene *myotonic dystrophy associated homeodomain protein* (*DMAHP*). These genes are both expressed in a wide range of tissues and their levels of expression have been shown to be altered in DM patients.

DMAHP is a member of the *Six* subfamily of genes of which 6 mammalian and 3 *Drosophila* genes have been identified. All members of this family encode a homeodomain that is highly diverged from other known homeodomains and a second homologous domain, the Six domain, which lies immediately upstream of the homeodomain. Homeodomain proteins are known to be transcription factors that control gene expression by binding to DNA and activating or repressing transcription. The work presented in this thesis investigated the DNA binding properties of the DMAHP protein.

Human foetal cDNA libraries were screened for the presence of a full length *DMAHP* cDNA but none was identified. Therefore, a cloned RT-PCR product and a genomic cosmid were used to subclone the Six domain, the homeodomain and both domains together into the bacterial expression vector pGEX4T3. From these clones, 3 glutathione S-transferase (GST)-DMAHP fusion proteins were produced and affinity purified by chromatography with a glutathione ligand. Affinity purified recombinant proteins and cell lysates from bacteria expressing the recombinant proteins were used in gel retardation assays with DNA oligonucleotides representing putative DNA binding sites. The putative sites included 2 in the promoter region of *DMPK* and the known DNA binding site of another Six subfamily member, AREC3/Six4 (the ARE). The recombinant fusion proteins showed no affinity for the 2 putative sites in *DMPK*. However, both the recombinant fusion proteins containing the homeodomain bound to the ARE, with the recombinant protein containing both domains (GST-Six+HD) forming 2 complexes, one of which was presumed to be a dimer complex. A cofactor present in the bacterial cell lysate increased the affinity of GST-Six+HD for the ARE.

Finally, a whole genome PCR based screen was used to identify genomic DNA sequences to which DMAHP binds, as an initial stage in the identification of genes regulated by DMAHP.

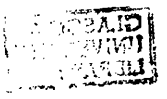

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Keith Johnson for allowing me to use his laboratory to perform the work presented in this thesis and for many hours of helpful and informative discussions. I would also like to thank all the members of his group, both those who are currently working in Glasgow and those that have now moved away, for the support and advice that they have given me over the last few years. I would particularly like to thank Cathy Boucher for the many techniques that she taught me during my first year and Catherine Winchester for the enormous amount of help she has given me with regard to the production and purification of recombinant proteins. The advice and suggestions that I received from Mark Bailey, Rohit Mistry and Graham Brock have been invaluable and I am extremely grateful to Graham Hamilton for solving my many computer problems. I am also very grateful for the DNA clones that I received from Cathy Boucher and the antibodies given to me by Catherine Winchester.

I would also like to thank Martin Boocock and Peggy Shelbourne for the long and informative discussions that helped to refine many of my experimental techniques. I am indebted to Keith Johnson, Catherine Winchester and Martin Boocock for the critical reading of this thesis.

This work was funded by the Muscular Dystrophy Group of Great Britain and Northern Ireland.

I wish to thank Kathy Williams and Teresa Fortune for the many mugs of hot chocolate and pints of beer that have helped to keep me sane throughout this project. My final thanks go to Adam and my parents who have supported me through the good times and the bad times.

To Mam, Dad and Adam

The research reported in this thesis is my own original work, except where otherwise stated and has not been submitted for any other degree.

Sarah E Harris

February 1999

List of Abbreviations

APS	ammonium persulphate
ARE	Na^+, K^+ ATPase α1 subunit gene regulatory element
AREC3	Na^+, K^+ ATPase α1 subunit gene regulatory element complex 3 (alternative name, SIX4)
ATPase	adenosine 5'-triphosphatase
ATP1A1	Na^+, K^+ ATPase α1 subunit gene
bp	base pair
BSA	bovine serum albumin
Ca^{2+}	calcium ion
CAG	triplet of cytosine, adenine and guanine
CAT	chloramphenicol acetyltransferase
CCG	triplet of cytosine, cytosine and guanine
cDNA	complementary deoxyribonucleic acid.
CIAP	calf intestinal alkaline phosphatase
cpm	counts per minute
cps	counts per second
CTG	triplet of cytosine, thymine and guanine
DM	myotonic dystrophy
DMAHP	myotonic dystrophy associated homeodomain protein (alternative name, SIX5)
DMPK	myotonic dystrophy protein kinase
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DRPLA	dentatorubral-pallidoluysian atrophy
DTT	dithiothreitol
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
dNTP	2'-deoxynucleoside 5'-triphosphate
dGTP	2'-deoxyguanosine 5'-triphosphate
dTTP	2'-deoxythymidine 5'-triphosphate
E	embryonic day
ECL	enhanced chemiluminescence
E.coli	*Escherichia coli*
EDTA	ethylenediaminetetracetic acid
EMBL	European Molecular Biology Laboratory

EST	expressed sequence tag
FA	Friedreich's ataxia
FRAXA	fragile X syndrome
FRAXE	fragile X site E (mental retardation)
GAA	triplet of guanine, adenine and adenine
GAL	galactosidase
GST	glutathione S-transferase
HD	Huntington's disease
HEPES	(N-[2-hydroxyethyl]piperazine-N'-2[2-ethanesulphonic acid])
HGMP	human genome mapping project
HPE	holoprosencephaly
HRP	horseradish peroxidase
IgG	immunoglobulin G
IPTG	isopropylthio-β-D-galactoside
K^+	potassium ion
kb	kilobase
kDa	kilo Dalton
LB	Luria-Bertani
LMP	low melting point
MEF3	myocyte-specific enhancer-binding factor 3
MCS	multiple cloning site
mRNA	messenger ribonucleic acid
Na^+	sodium ion
NCBI	National Centre for Biotechnology Information
NMR	nuclear magnetic resonance
OD	optical density
OLB	oligonucleotide labelling buffer
PBS	phosphate buffered saline
PBS-T	phosphate buffered saline-Tween®20
PCR	polymerase chain reaction
PEG	polyethylene glycol
PMSF	phenylmethylsulphonyl fluoride
RNA	ribonucleic acid
RNase	ribonuclease
RT-PCR	reverse transcriptase polymerase chain reaction
SAPU	Scottish Antibody Production Unit
SBMA	spinobulbar muscular atrophy
SCA	spinocerebellar ataxia

SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
Six	*sine oculis related homeobox*
so	*sine oculis*
TEMED	NNN'N'-tetramethlethylene-diamine
Tris	Tris(hydroxymethyl)amino methane
UTR	untranslated region
UV	ultra violet
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside

Note: To avoid confusion with members of the *Six* subfamily of genes, except at the beginning of sentences all numbers have been written as arabic numerals.

Chapter 1 Introduction

1.1 Myotonic Dystrophy

Myotonic dystrophy, otherwise known as dystrophia myotonica (DM), was first described in 1909 as a disorder causing atrophy and weakness of the face and distal limb muscles and myotonia (Batten and Gibb, 1909; Steinert, 1909). DM is inherited as an autosomal dominant disorder and is the most frequently occurring form of adult muscular dystrophy, with a world wide incidence of 1 in 8,000 (Harper, 1989). Clinically the disease is now known to be highly variable with symptoms including; myotonia, progressive weakness and wasting of the muscles, cardio-respiratory problems, cataracts, gastro-intestinal problems, hyperinsulinism, hypersomnia, premature balding, mental retardation, gonadal atrophy and male sterility. Patients with the mildest form of the disease develop symptoms late in life and cataracts are often the only detectable sign. In the early stages of the classical form of DM, weakness of the facial, jaw and sternomastoid muscles are the earliest muscle symptoms followed by distal muscle weakness. This weakness is caused mainly by the atrophy of type I muscle fibres and proximal muscles become affected as the disease progresses. DM patient skeletal muscle biopsies show variation in muscle fibre size and shape and an increased number of split muscle fibres and connective tissue. An increased number of central nuclei are seen in the muscle fibres, which is presumably due to regeneration as a result of the atrophy.

Although the disease generally presents itself in the third or fourth decades, it can occur congenitally and in these cases the symptoms are usually more severe and include hypotonia, facial displegia and weakness of sucking and swallowing due to delayed muscle maturation. Congenital DM is inherited almost exclusively from DM-affected mothers and has a high mortality rate. Survivors develop the classic form of the disease at about 10 years of age (Harper, 1989).

1.2 The Pathophysiological Mechanism of DM

Several attempts have been made to establish the mechanism by which the properties of the diseased systems are altered. However, due to the use of denervated and aneurally grown skeletal muscle fibres (which are very different from adult tissue) in many of the studies, conflicting electrophysiological results have been reported and the complete mechanism is still unknown.

The protein kinase activity in erythrocyte membranes of DM patients was shown to be reduced compared to normal controls indicating a membrane abnormality in DM (Roses and Appel, 1973). Therefore, the activity of the Na^+, K^+ -ATPase pump,

which is required for the removal of intracellular Na^+ after the transmission of an action potential and Na^+ concentration have been investigated. A decrease in the number of Na^+ pumps have been reported (Desnuelle *et al.*, 1982) along with increased levels of Na^+ ions in skeletal muscle fibres of DM patients (Edstrom and Wroblewski, 1989; Rudel and Lehmannhorn, 1985). However, both a decrease in Na^+, K^+ -ATPase activity (Benders *et al.*, 1996; Benders *et al.*, 1993) and no change in activity have been reported (Roses and Appel, 1974).

Other ion channels have also been implicated in the pathology of DM. Apamin (a peptide from bee venom that blocks small conductance Ca^{2+}-activated K^+ channels) receptors are present in DM muscle but absent in normal adult control muscle (Renaud *et al.*, 1986). The possible role of these apamin sensitive K^+ channels in the genesis of myotonia was investigated by the injection of apamin into the thenar muscles of DM patients. This decreased their basal electrical activity and reduced the intensity and duration of myotonic discharges suggesting that these K^+ channels are involved in the mechanism of myotonia (Behrens *et al.*, 1994). However, other embryonic skeletal muscle markers, *neural cell adhesion molecule* (*N-CAM*) and *5.1H11*, the regulation of which is coordinate with apamin binding, are both expressed in DM muscle (Walsh *et al.*, 1988). Therefore, the expression of these 3 embryonic genes in DM patient muscle may be a reaction to the pathology of the disease rather than a direct cause of the symptoms.

A recent study used the single-channel patch-clamp technique to investigate Ca^{2+}-activated K^+ channels of erythrocytes from patients with DM (Pellegrino *et al.*, 1998). The basic properties of single Ca^{2+}-activated K^+ channels; conductance, rectification, kinetics, voltage and calcium dependence were indistinguishable from normal controls. However, the activity of channels from DM patient erythrocytes exhibited mean patch currents that were significantly higher than the controls, due to a higher opening frequency, associated with a reduced mean channel closed time. Therefore, Ca^{2+}-activated K^+ channels of erythrocytes from DM patients either detect a higher intracellular Ca^{2+} concentration and/or express an increased Ca^{2+} sensitivity. These experiments identified defective kinase mechanisms at the molecular level.

1.3 The Myotonic Dystrophy Mutation

The primary genetic defect responsible for myotonic dystrophy was identified as an expanded trinucleotide $(CTG)_n$ repeat in the 3' untranslated region (UTR) of a protein kinase gene, *myotonic dystrophy protein kinase* (*DMPK*) on chromosome 19 (Aslanidis *et al.*, 1992; Brook *et al.*, 1992; Buxton *et al.*, 1992; Fu *et al.*, 1992; Harley *et al.*, 1992; Mahadevan *et al.*, 1992). Sequence analysis showed DMPK to

have homology with members of the serine/threonine family of kinases. Generally, the greater the number of repeats that a patient has, the more severely affected they are (table 1.1). Normal individuals have between 5 and 37 repeats, mildly affected patients have 50-150 repeats, adult onset (classical) patients 100-1,000 repeats and congenital cases can have more than 2,000 repeats. In general there is an earlier onset of DM as well as an increase in severity of the clinical symptoms with transmission to successive generations within families (Harper *et al.*, 1992). This genetic anticipation is accompanied by an increase in the number of $(CTG)_n$ repeats (Mahadevan *et al.*, 1992).

Phenotype	**Clinical symptoms**	**Repeat length**
MILD	Cataracts	50-150
CLASSICAL	Myotonia, Muscle wasting, Premature balding, Gonadal atrophy, Cardiac conduction defects	100-1000
CONGENITAL	Hypotonia, Mental Retardation, Facial displegia	>1000

Table 1.1 Correlates between phenotype and repeat expansion lengths

The meiotic instability of the DM mutation has also been shown to be dependent on the sex of the parent as well as the size of the parental repeat. For $(CTG)_n$ sequences of less than 0.5 kb a positive correlation between size of repeat and intergenerational enlargement was found almost equally in both male and female meioses. With repeats greater than 0.5 kb intergenerational variation is more important in female meioses and contraction of repeat length occurs almost exclusively in male meioses for fragments greater than 1.5 kb (Lavedan *et al.*, 1993). This may, in part, be due to a selection barrier during spermatogenesis preventing the survival of sperm with long repeats. Congenital DM, which presents the most severe phenotype was thought to be exclusively maternally inherited (Hofmann-Radvanyi *et al.*, 1993). However, cases of paternally inherited congenital DM have been described, which shows that it is possible for transmission of a large allele by a male to occur although it is a very rare event. For an example see Nakagawa *et al.* (1994). This excludes maternal imprinting, mitochondrial inheritance or transplacental factors as being of importance in congenital DM.

In addition to germline instability, somatic heterogeneity between and within tissues has been demonstrated. Repeat lengths were found to be greater in muscle than in lymphocytes from the same patient (Anvret *et al.*, 1993; Thornton *et al.*, 1994). As muscle is the tissue that is primarily affected, these results suggest that for presymptomatic testing, muscle DNA rather than lymphocyte DNA should be tested. The occurrence of mitotic repeat length variation within DM patients was assessed by analysing a broad range of tissues from single patients and monozygotic twins (Jansen *et al.*, 1994). Differences in the repeat length were seen in all tissues from single DM patients and the twins except when the patients carried extremely long repeats. The repeat lengths in fathers' sperm were also compared to their offspring's blood and were found to be significantly different. It was concluded that DM patients are both somatic and germline tissue mosaics and that $(CTG)_n$ expansions occur during early embryonic mitotic divisions in both somatic and germline tissue formation. Somatic variation within tissues has also been demonstrated by small pool PCR (Monckton *et al.*, 1995). Using standard techniques that analyse simultaneously at least 10^4 cells individual alleles cannot be distinguished and a heterogeneous smear of $(CTG)_n$ repeats is observed when visualising expanded alleles from DM patients. Small pool PCR demonstrated that the smear is composed of multiple coalesced fragments that can be resolved into discrete length alleles derived from single cells. It was also demonstrated that somatic expansions in the parent can contribute to apparent intergenerational repeat length differences. It is these somatic expansions that lead to different length repeats in different cells of the same tissue. It has also been shown that the repeat length within individual patients increases with age, showing ongoing mitotic instability in DM (Martorell *et al.*, 1995). The degree of expansion correlates with the initial size of the repeat.

1.4 Origin of the myotonic dystrophy mutation

The DM $(CTG)_n$ repeat is polymorphic in the general population, ranging from 5-37 repeats (Brunner *et al.*, 1992). In African, Asian, European, Japanese and Native American populations a trimodal distribution of these normal sized repeats has been reported. $(CTG)_5$ repeats form the first mode and are the most frequently occurring in European populations (Brunner *et al.*, 1992; Davies *et al.*, 1992). The second mode occurs most frequently in non-European populations and consists of repeats ranging in size from 11-15 (Davies *et al.*, 1992; Watkins *et al.*, 1995; Zerylnick *et al.*, 1995). Alleles with repeats greater than 19 form the third mode (Watkins *et al.*, 1995; Zerylnick *et al.*, 1995).

The mutation responsible for DM has been found to be in complete linkage disequilibrium with an *Alu* element insertion/deletion (+/-) polymorphism. This suggests that DM stems from one, or possibly a few, ancestral mutations (Mahadevan *et al.*, 1993b). Imbert *et al.* (1993) proposed a mulitistep model of evolution for the repeat and the origin of DM mutations. The linkage disequilibrium between normal alleles at the repeat and the nearby *Alu* (+/-) element was analysed and the major $(CTG)_5$ allele and the large alleles $(CTG)_{>19}$ were found to show the same linkage disequilibrium as the DM mutation chromosomes. They are all in linkage disequilibrium with the *Alu* (+) allele and alleles with 11-13 repeats are associated with the *Alu* (-) allele. This suggests that chromosomes with more than 19 repeats are derived from the $(CTG)_5$ allele by a small number of events. The DM mutation chromosomes are then derived from a pool of alleles with 20-30 repeats. High resolution genetic analysis of the DM locus using PCR based assays of 9 polymorphisms was performed and four main haplotypes were observed (A-D) (Neville *et al.*, 1994). Haplotype A is *Alu* (+) whereas haplotypes B-D are *Alu* (-). DM is in complete association with haplotype A, the most common haplotype in the normal population. Haplotype A in the normal population is associated with 5-37 $(CTG)_n$ repeats whereas haplotypes B-D are associated with only 11-14 repeats. Again these results indicate that $(CTG)_{20-30}$ alleles on haplotype A act as ancestors for alleles with larger disease causing repeats. Neville *et al.* (1994) therefore proposed that the DM mutation occurred on the background of a particular haplotype (A) in which the $(CTG)_n$ repeat became unstable and predisposed to amplification. To test this hypothesis the $(CTG)_n$ repeat distributions of 2 populations with a low incidence of DM and 2 populations with a high incidence of DM were examined (Rubinsztein *et al.*, 1994). It was determined that populations with a low frequency of DM have fewer alleles with greater than 19 repeats, suggesting that large normal alleles are indeed a reservoir for the creation of new DM mutants. No repeats of less than 5 were detected and it was suggested that this may be due to selection against smaller alleles or that chromosomes with 5 repeats may mutate to larger expansions most of the time, if they mutate at all.

Krahe *et al.* (1995b) conducted a detailed haplotype analysis of the DM region of a Nigerian DM family. Affected members of the family each had one DM allele with an expanded $(CTG)_n$ repeat. However, there was dissociation of the allele from other alleles of the putative predisposing haplotype suggesting that the repeat expansion in this family was caused by an independent mutational event and thus refutes the hypothesis that a single haplotype predisposes to repeat expansion. However, this study supports the theory that an expanded $(CTG)_n$ repeat is required for DM to occur.

The association between normal alleles at the $(CTG)_n$ repeat and two nearby polymorphisms (the *Alu* and (*DMPK*) (G/T) intron 9/*Hinf* I polymorphisms) in South African negroids has been analysed (Goldman *et al.*, 1995). No cases of DM have ever been described in this population and repeats > 19 are extremely rare. As well as finding the haplotypes seen in Caucasoids and Japanese other haplotypes not previously seen were observed. Goldman *et al.* (1995) suggest that only a small number of these "African" chromosomes were present in the progenitors of non-African peoples and that the rare mutations occurred after the migration from Africa.

The fact that DM, a disease which normally decreases reproductive fitness within a few generations, has been maintained in the population is very interesting. Carey *et al.* (1994) report on the distortion of the ratio of DM to normal alleles being transmitted to the next generation. They suggest that healthy male individuals heterozygous for DM alleles, in the normal size range, preferentially pass on alleles of greater than 19 repeats, allowing potential DM mutations to be replenished in the population. A second group have also concluded that meiotic drive is responsible for maintaining the level of DM alleles in the population (Gennarelli *et al.*, 1994). However, it has been suggested that these results were not statistically valid and that the observed segregation distortion may be an artefact of testing multiple hypotheses from the same data (Hurst *et al.*, 1995). A simpler analysis that scored the alleles of an individual as either long or short and did not specifically look at alleles with greater than 19 repeats concluded that there is preferential transmission of the longer allele during female meiosis but not during male meiosis (Chakraborty *et al.*, 1996). In support of these results a study that analysed single sperm from 3 males who were heterozygous at the DM locus and had one allele larger than and one allele smaller than 19 repeats found no evidence of meiotic segregation distortion (Leeflang *et al.*, 1996).

1.5 Effects of $(CTG)_n$ Repeat

At the time of starting this project 7 other diseases were known to be caused by trinucleotide repeat instability; Kennedy's disease (spinal and bulbar muscular atrophy, SBMA), fragile X syndrome (FRAXA), Huntington's disease (HD), spinocerebellar ataxia, types 1 and 3 (SCA1 and SCA3), FRAXE mental retardation and Haw-River Syndrome (dentatorubral pallidoluysian atrophy, DRPLA) (Kawaguchi *et al.*, 1994; Knight *et al.*, 1993; Koide *et al.*, 1994; La Spada *et al.*, 1991; MacDonald *et al.*, 1993; Orr *et al.*, 1993; Verkerk *et al.*, 1991). Friedreich's ataxia (FA), SCA2, SCA6 and SCA7 have since been added to this list (Campuzano *et al.*, 1996; David *et al.*, 1997; Imbert *et al.*, 1996; Zhuchenko *et al.*, 1997). In all these examples different pathophysiological mechanisms to those underlying DM have been invoked. FRAXA, FRAXE (CCG repeats) and FA (GAA repeat) are due to diminished

expression of their gene products and are recessive in nature (Campuzano *et al.*, 1996; Pieretti *et al.*, 1991). The other disorders arise through expansions of CAG trinucleotides encoding polyglutamine tracts in translated exonic regions of their genes, with an associated gain of function, potentially as a result of altered protein conformations (Trottier *et al.*, 1995). Given the location of the repeat in the 3'-UTR of *DMPK* and the autosomally dominant inheritance, it seems likely that a different mechanism underlies the pathophysiology associated with DM.

1.5.1 DNA Effects

Expanded CTG repeats have been shown to form the strongest known natural nucleosome positioning elements (Wang and Griffith, 1995). The efficiency of the nucleosome formation was seen to increase with the size of the repeat. This suggests that expanded CTG blocks alter local chromatin structure which in turn affects the expression of *DMPK* and other neighbouring genes. Further evidence for repeat expansion altering chromatin structure came from studies in muscle nuclei of 3 unrelated DM patients with large expansions. They were shown to exhibit DNase I resistance and inaccessibility to nucleases on the expanded allele at the normally hypersensitive site adjacent to the repeat (Otten and Tapscott, 1995). Stabilisation of the nucleosome structure may account for the loss of DNase I hypersensitivity with concomitant repression of transcription through the region. However, the loss of hypersensitive sites may be due to the binding of other protein factors to the region as a result of repeat expansion.

The possibility of *DMPK* being an imprinted gene is unlikely given the lack of methylation differences of CpG islands within *DMPK* between paternally and maternally derived alleles in congenital and adult onset patients (Shaw *et al.*, 1993a). However, a recent report indicates that at least in congenital patients, increased methylation occurs on the 5' side of the $(CTG)_n$ repeat on the DM allele and that this increased methylation leads to a decrease in binding of Sp1 to a consensus binding site (Steinbach *et al.*, 1998).

1.5.2 RNA and Protein Effects

As the expanded triplet repeat occurs in the 3'-UTR of the *DMPK* gene it cannot directly affect protein sequence. However, several possible mechanisms by which it affects *DMPK* gene expression have been suggested (Whiting *et al.*, 1995). These include:

(a) alteration in steady-state levels of *DMPK* expression.

(b) interference with normal splicing patterns for *DMPK* RNA.

(c) interruption of normal regulatory role of *DMPK* transcripts.

(d) change in efficiency of translation.

When this project was initiated, conflicting data concerning the effect of the mutation on mRNA and protein levels had been published. Tissues from adult myotonic dystrophy patients were analysed by quantitative RT-PCR and radioimmunassay and this showed decreased levels of mRNA and protein expression implying that reduction in *DMPK* leads to the disease by disruption of signal transduction and amplification pathways (Fu *et al.*, 1993). Carango *et al.* (1993) also reported reduction in both the synthesis and the processing of *DMPK* mRNA. However, Sabourin *et al.* (1993) showed that *DMPK* steady state levels were markedly increased in congenital DM tissues and it was demonstrated that increases are due to elevated levels of transcripts derived from mutant *DMPK* alleles. These conflicting results may be due to differences in the tissues studied. Adult tissues were used by 2 groups (Carango *et al.*, 1993; Fu *et al.*, 1993) whereas cells and tissues from severely affected neonates were used by the third (Sabourin *et al.*, 1993). Congenital cases are known to have different symptoms to adult onset patients, for example, adults tend to be myotonic whereas infants tend to be hypotonic. The control tissues also differed between the two groups, one using tissues from patients with neurologic diseases (Fu *et al.*, 1993) and the another using mRNAs from healthy adults and foetal tissues (Sabourin *et al.*, 1993). Different molecular techniques were used in the studies, adding to the difficulties in directly comparing the results.

More recently, the allele-specific expression of *DMPK* has been investigated and shown not to be affected at the transcriptional level, but rather at the processing and maturation level (Krahe *et al.*, 1995a). The accumulation of transcripts from both normal and expanded alleles in patient muscle were compared with normal and myopathic controls (Wang *et al.*, 1995). Only a small decrease of *DMPK* RNA in the total RNA pool from muscle was found and this was not disease specific, as the myopathic controls showed a similar decrease. However, dramatic decreases of both mutant and normal *DMPK* poly $(A)^+$ RNA were observed and this was disease specific. The authors speculate that normal and expanded *DMPK* genes are transcribed in patient muscle, but RNA with an abnormal expansion has a dominant effect on RNA metabolism preventing the accumulation of poly$(A)^+$ RNA. The intracellular localisation of transcripts from the *DMPK* gene in fibroblasts and muscle biopsies from DM patients and normal controls was determined using a fluorochrome-conjugated probe which hybridised specifically to the $(CTG)_n$ repeat expansion. No significant difference in the cytoplasmic location of the *DMPK* mRNA was observed (Taneja *et al.*, 1995). However, the probe detected transcripts as bright foci in the nuclei of DM patient fibroblasts and muscle but not in samples from normal controls.

Using a probe that hybridised to the 5' end of the *DMPK* gene, and would therefore detect all *DMPK* sequences, as well as the expansion specific probe it was shown that focal accumulation of post transcriptional RNA only occurs with transcripts from the affected allele. In the steady-state, nuclear export from the affected allele does not appear to be significantly suppressed, as RNA containing the repeat is present in the cytoplasm at a level similar to that of normal RNA. However, post transcriptional foci may inhibit other nuclear functions indirectly. It is now believed that the earlier contradictory reports were due to the use of different RNA extraction methods. Some methods remove only cytoplasmic DNA which results in a decrease in the number of *DMPK* transcripts recovered, whereas if both nuclear and cytoplasmic RNA are analysed an increase in the number of *DMPK* transcripts is detected, due to the their accumulation in the nuclei. More recent theories on the effect of the $(CTG)_n$ repeat expansion at the RNA level, including the identification of CUG binding proteins are discussed in section 7.6.2.

Biochemical and histochemical studies have suggested that a 53 kDa *DMPK* product is localised at sites of cell-cell contacts in neuromuscular tissues, in particular at neuromuscular and myotendinous junctions of human and rodent skeletal muscles and intercalated disks in cardiac muscle. This indicates that *DMPK* probably plays a specialised role in intercellular communication in striated muscle. It is possible that defects in communication could lead to muscle atrophy, myotonia and heart failure (Van der Ven *et al.*, 1993). A similar sized protein (54 kDa) that is recognised by DMPK antiserum has recently been characterised and shown to display a serine/threonine kinase activity in the heart and tyrosine kinase activity in the skeletal muscle (Etongue-Mayer *et al.*, 1998). As the predicted molecular weight of DMPK is approximately 70 kDa it is possible that this protein is a cross-reacting kinase species and unrelated to DMPK. It has also been demonstrated that *DMPK* knockout mice still express this 54 kDa protein which supports the theory that this protein is not DMPK (Jansen *et al.*, 1996). A monoclonal antibody has now been used to identify a 64 kDa protein in type 1 skeletal muscle localised to the triadic region, and a 79 kDa protein in brain (Dunne *et al.*, 1996). Comparisons of proteolytic digests of the muscle protein and recombinant DMPK show similar mobilities, indicating that the 64 kDa protein is likely to be a *bonafide* isoform. In severely affected DM patients redistribution of this isoform to the peripheral sarcoplasmic masses, consistent with previously reported altered localisation and expression of transcripts has been seen (Sabourin *et al.*, 1993).

A new panel of monoclonal antibodies has recently been used in an attempt to determine the true size of the DMPK protein and its localisation within tissues (Pham *et al.*, 1998). This study identified 2 isoforms of DMPK that have molecular weights of 72 and 80 kDa. The 72 kDa isoform but not the 80 kDa isoform was detected in all tissues tested. Previous studies have also detected doublets of approximately these molecular weights (Maeda *et al.*, 1995; Whiting *et al.*, 1995). The presence of DMPK in the intercalated disks of cardiac muscle but

not in the neuromuscular junctions of skeletal muscle was confirmed. A 55 kDa protein was detected in skeletal muscle by 2 of the 12 monoclonal antibodies but it was not thought to be an alternatively spliced form of DMPK as it was not detected by antibodies raised against the coil domain. To date no isoforms of DMPK lacking the coil domain have been reported (Fu *et al.*, 1993; Jansen *et al.*, 1992; Mahadevan *et al.*, 1993a).

1.6 Other Genes in the Region of the $(CTG)_n$ Repeat

The complexity of the clinical phenotype of DM along with the initial contradictory data concerning the effects of $(CTG)_n$ repeat expansion on the expression of *DMPK* and the alteration in chromatin structure surrounding the repeat, led researchers to look for genes neighbouring *DMPK*.

1.6.1 *DM-N9 (59 gene)*

DM-N9 is a murine gene that is located, with its 3' end, 1.1 kb upstream of the *Dmpk* gene (Jansen *et al.*, 1995; Jansen *et al.*, 1992). This 7 kb gene contains 5 exons and codes for a 650 amino acid protein. Two regions within this gene show significant homology to WD repeats, which are highly conserved amino acid sequences found in a family of proteins involved in signal transduction of cell regulatory functions, although the actual function of *DM-N9* is, as yet, unknown. *DM-N9* is expressed strongly in brain and testis and could therefore be a candidate for involvement in mental and testicular symptoms in severe cases of DM. The human homologue to *DM-N9* (*59*) has been identified and shown to be expressed in heart, brain, liver, kidney, spleen and testis (Shaw *et al.*, 1993b).

1.6.2 *DMAHP/Six5*

The presence of an extensive CpG island at the 3' end of *DMPK* indicated that a novel gene lay immediately downstream of the $(CTG)_n$ repeat. This region was sequenced and murine and human sequences were found to share 86% identity at the putative amino acid level (Boucher *et al.*, 1995). The $(CTG)_n$ repeat was shown to be within a 3.5 kb CpG island. Sequence data indicated that the gene encoded a homeodomain protein and was therefore called *DM locus-associated homeodomain protein* (*DMAHP*). *DMAHP* has now been identified as being one of a growing number of members of the *sine oculis related homeobox* (*Six*) subfamily of genes (section 1.8), so called because the original member to be identified was the *Drosophila* eye development gene *sine oculis* (*so*). Therefore, when a murine cDNA *DMAHP*

homologue was identified it was named *Six5* (Kawakami *et al.*, 1996b). All members of the *Six* subfamily share a region of a high sequence homology which encodes a 60 amino acid homeodomain and immediately upstream of the homeodomain a domain unique to the *Six* subfamily, the Six domain which is approximately 116 amino acids. RT-PCR on mRNA preparations showed that *DMAHP* is expressed in many tissues including skeletal muscle, fibroblasts, lymphocytes, heart and brain in both normal adults and DM patients. An enhancer element that controls the expression of *DMAHP* was recently identified within the DNase I hypersensitive site adjacent to the $(CTG)_n$ repeat (section 1.5.1). A 2 to 4 fold reduction in the steady state transcript levels of *DMAHP* from the expanded allele compared to the normal allele of DM patient's, was seen in fibroblasts and skeletal muscle cells implying that expansion of the $(CTG)_n$ repeat does indeed alter local chromatin structure in such a way as to reduce expression of *DMAHP*, which in turn is involved in the pathogenesis of DM (Klesert *et al.*, 1997). A second group have also reported a reduction in the levels of *DMAHP* mRNA synthesis from the DM allele in primary myoblast cultures and postmortem muscle, heart and brain tissues from DM patients (Thornton *et al.*, 1997). However, a contradictory report indicated that levels of *DMAHP* expression were unchanged in cell lines from DM patients (Hamshere *et al.*, 1997). Gene expression patterns can differ between tissues and myoblast cultures and therefore gene expression data from cell lines may not be representative of what occurs in human tissue. This is partially due to the lack of innervation in cell lines. However, the dynamics of the $(CTG)_n$ repeat are also altered in cultured lymphoblastoid cell lines from DM patients and this may affect the expression of genes that are altered by the repeat expansion (Ashizawa *et al.*, 1996). The structure of the DM locus indicating *DMAHP*, *DMPK* and *59* is illustrated in figure 1.1.

1.6.2.1 Genomic Structure of *DMAHP*

Two alternative transcripts of human *DMAHP* were identified by RT-PCR (Boucher *et al.*, 1995). One isoform contains exons A, B and C and the second isoform contains only exons A and C and has a different reading frame in exon C (figure 1.2). The Six box and homeobox are situated in exon A and therefore encoded by both isoforms. To date only the A-B-C isoform has been identified in mouse (Heath *et al.*, 1997). The A-B-C isoform is predicted to encode a protein of about 82 kDa and the A-C isoform a protein of about 36 kDa. *DMAHP* lacks a consensus TATA box upstream of the open reading frame and consistent with this finding multiple transcription initiation sites have been identified in both the human and mouse genes (Klesert *et al.*, 1997; Murakami *et al.*, 1998). The human transcriptional start sites cover a 50-70 bp

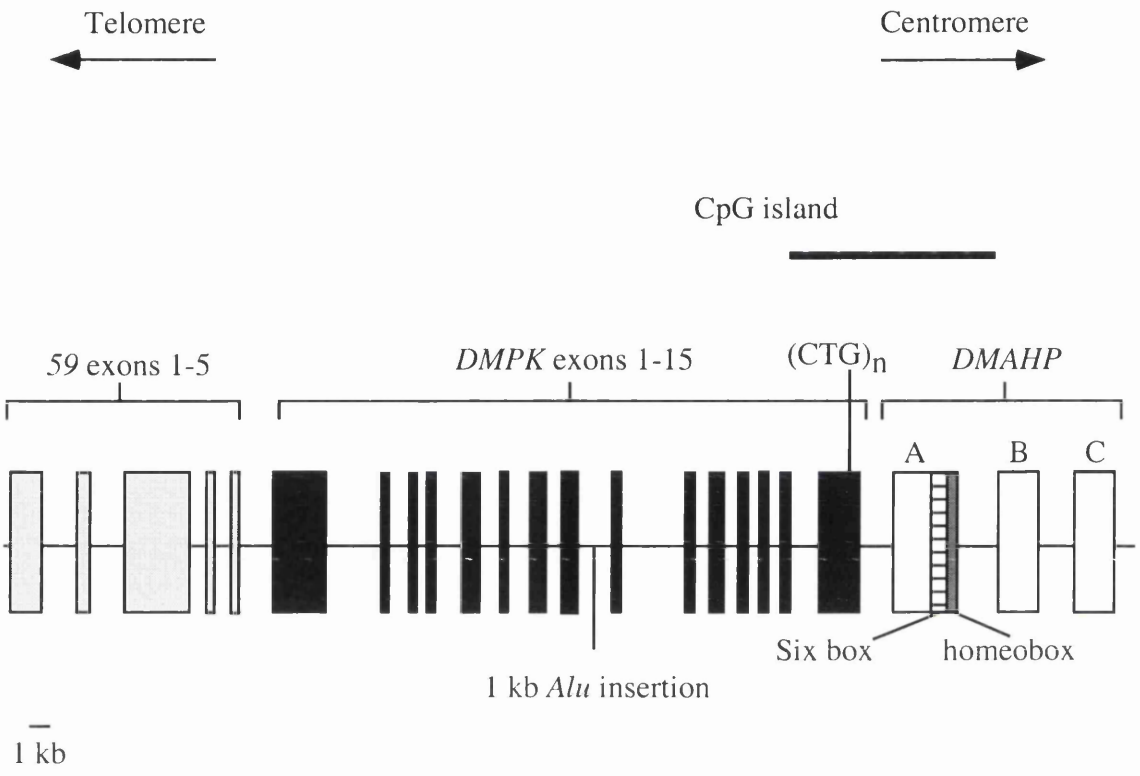

Figure 1.1 Genomic structure of the DM region

A schematic diagram of the DM region showing the relative positions of *59*, *DMPK* and *DMAHP*. The extensive CpG island, the $(CTG)_n$ repeat and the 1 kb *Alu* insertion/deletion polymorphism are indicated.

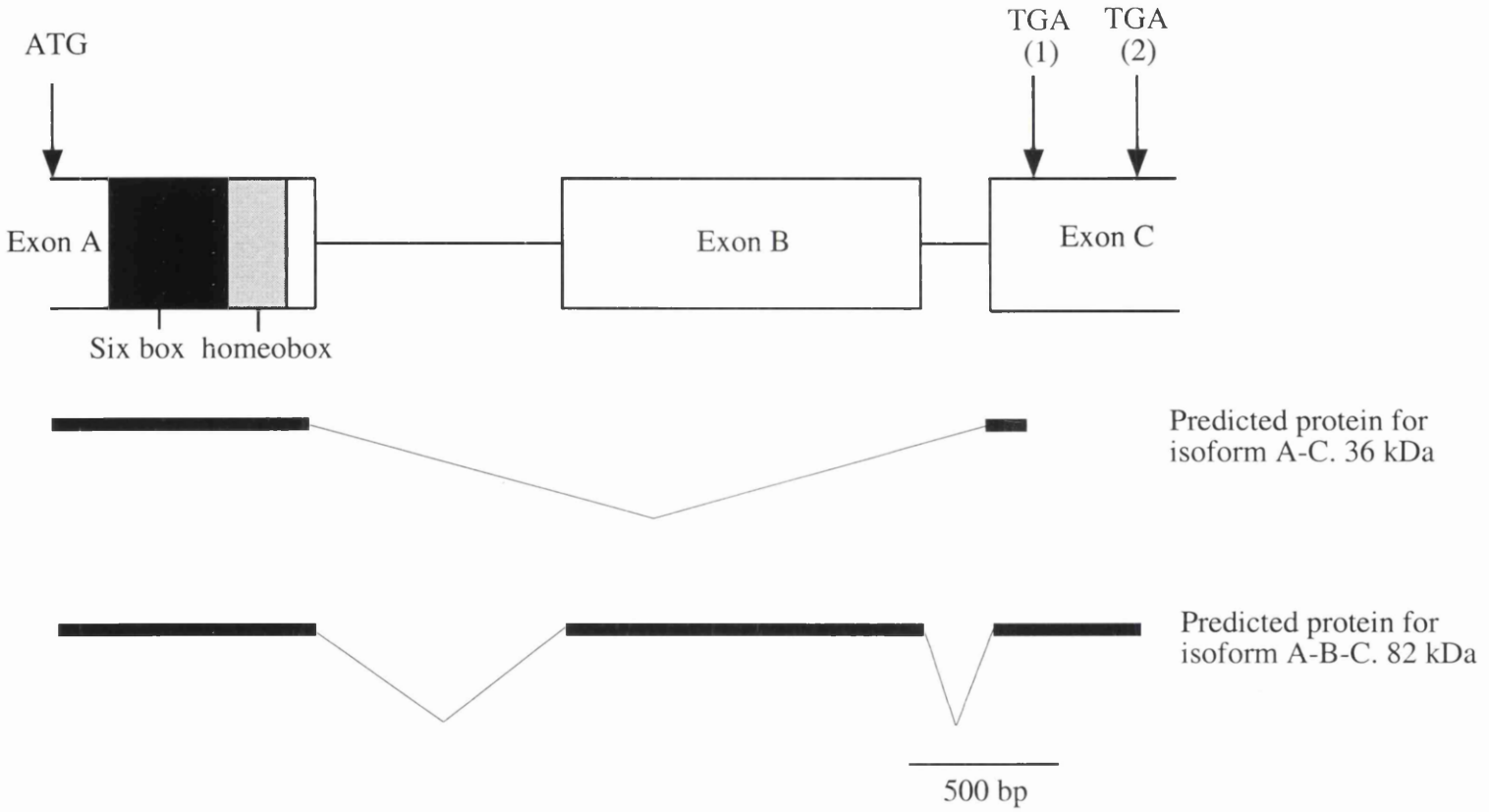

Figure 1.2 Structure of *DMAHP* and predicted proteins

A schematic diagram of *DMAHP* showing the structures of the 2 predicted protein isoforms. The predicted stop sites for the A-C isoform (1) and the A-B-C isoform (2) are shown.

region and are all within 400 bp of the coding region and the mouse sites are in an equivalent position in the mouse gene. Transient transfection assays were used to identify regulatory regions in the mouse gene promoter region. Positive elements were identified between -434 and -344, -317 and-270 and -137 and -97 and negative elements were identified between -344 and -317 and -224 and -178. Gel retardation assays using nuclear extracts from P19 cells along with super-shift assays using Sp1 and Sp3 antibodies confirmed that the 3 positive regulatory regions and 1 of the negative regulatory regions contained Sp1/Sp3 binding elements. These Sp1/Sp3 binding sites are conserved between mouse and human and the enhancer element in the human gene (section 1.6.2) was also shown to be active in mouse (Murakami *et al.*, 1998).

1.7 Homeobox Genes

Transcription factors are proteins that bind to DNA in a sequence-specific manner and interact with other components of the transcriptional machinery to enhance or inhibit transcription of specific target genes. The majority of transcription factors contain a DNA binding domain and a transcriptional activation domain (Engelkamp and van Heyningen, 1996). Homeobox genes encode a conserved domain of about 60 amino acids (the homeodomain) that can bind to DNA and a transcriptional activation/repression domain, thus acting as transcription factors. Recently, the *Drosophila* homeodomain protein bicoid, that transcriptionally activates target genes by binding to DNA has been shown to act as a repressor of translation by binding to the 3'-UTR of the *caudal* mRNA, thus blocking the initiation of caudal translation (Dubnau and Struhl, 1996). Therefore, homeodomain proteins can bind to both DNA and RNA.

1.7.1 Structure of the Homeodomain

The structures of several homeodomain proteins have been determined using nuclear magnetic resonance (NMR) spectroscopy. The homeodomain of the *Drosophila* Antennapedia homeodomain was expressed in a T7 expression vector in *E.coli* and purified to homogeneity (Muller *et al.*, 1988). NMR spectroscopy of free Antennapedia homeodomain in solution demonstrated that the homeodomain consists of a short N-terminal arm of 6 amino acids, and 4 helices (Billeter *et al.*, 1993; Otting *et al.*, 1988; Qian *et al.*, 1989). Helices I and II are arranged antiparallel to each other and are connected by a hexapeptide loop. Helices II and III are connected by a tripeptide, that forms a β-turn, constituting a helix-turn-helix motif and helix IV

immediately follows helix III, separated only by a slight bend. Together, helices III and IV are aligned perpendicular to the first two helices. The homeodomain has a compact globular structure from which helix IV protrudes and helices I, II and III are held together by a core of 11 amino acids (figure 1.3). Later NMR studies of *Drosophila* engrailed and the yeast mating type determinant MATα2 homeodomain proteins, in their DNA complexes, indicated that they too have a similar structure with the exception that the third and fourth helices form a single helical region (Kissinger *et al.*, 1990; Wolberger *et al.*, 1991).

1.7.2 DNA Binding Specificity of the Homeodomain

Based on comparisons with bacteriophage repressor proteins, the second α-helix of the helix-turn-helix motif (helix III of the homeodomain) was hypothesised to be the recognition helix, responsible for determining the DNA binding specificity of homeodomain proteins. Mutagenic analysis was used to determine which amino acids of this putative recognition helix are used to discriminate between different DNA sequences (Hanes and Brent, 1989). When the recognition helix of bicoid was exchanged for that of Antennapedia the protein had the binding specificity of Antennapedia. Further analysis determined that exchanging just the ninth amino acid in the recognition helix (amino acid 50 of the homeodomain) resulted in an identical change in binding specificity. DNA-protein binding experiments using the *Drosophila* homeodomain protein paired containing point mutations in the homeodomain also lead to the conclusion that amino acid 50 of the homeodomain is responsible for much of the protein's binding specificity (Treisman *et al.*, 1989). These experiments demonstrated the importance of amino acid 50 in determining DNA binding specificity.

NMR spectroscopy showed that Antennapedia binds to a 14 bp duplex as a monomer (Otting *et al.*, 1990). Helices III and IV lie in the major groove of the DNA and the N-terminal arm makes specific contacts with bases in the minor groove and the loop between helices I and II interacts with the DNA back bone on the other side of the major groove. *In vitro* analysis has been utilised to identify the binding specificity of many homeodomain proteins (table 1.2). The majority of the preferred binding sites contain the core sequence ATTA. Bases of this core sequence have been shown to interact, in the major groove, with amino acids 47 (isoleucine) and 51 (asparagine), which are in the recognition helix, and in the minor groove with amino acids 3 (arginine) and 5 (arginine), which are in the N-terminal arm, of Antennapedia and engrailed homeodomains (Billeter *et al.*, 1993; Kissinger *et al.*, 1990; Otting *et al.*, 1990). Yeast MATα2 differs from *Drosophila* Antennapedia and engrailed at amino

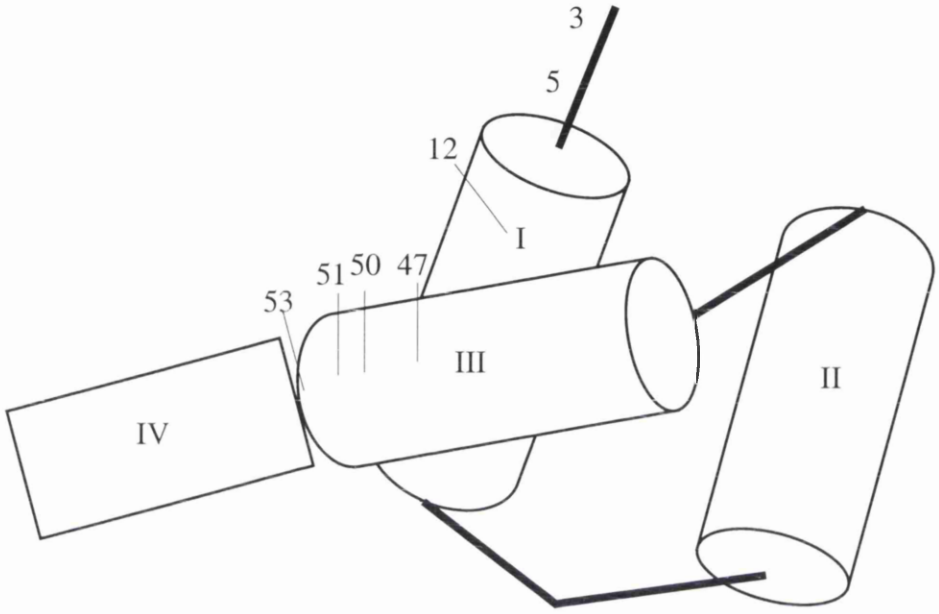

Figure 1.3 Structure of the Antennapedia homeodomain as determined by NMR spectroscopy

A schematic diagram of the homeodomain of the *Drosophila* Antennapedia protein. The α-helical regions (I-IV) are represented as cylinders. The positions of the amino acids mentioned in the text are indicated. Adapted from Latchman (1998).

acid positions 3, 5 and 47 and binds to the sequence GTAA rather than the classical core sequence which emphasises the importance of these amino acids with regard to binding specificity (Wolberger *et al.*, 1991). For homeodomain proteins that bind to the ATTA core sequence the amino acid at position 50 of the homeodomain (discussed above) is important in defining the 2 bases prior to the ATTA core. Homeodomain proteins that bind to GGATTA with a high affinity contain a lysine at position 50, for example bicoid and fushi tarazu. A consensus protein sequence has been derived from a compilation of 346 homeodomain sequences. Seven positions in this sequence are conserved in more than 95% of the 346 sequences, these include amino acids: 5 (arginine), 51 (asparagine) and 53 (arginine) that are directly involved in DNA binding (Gehring *et al.*, 1994). The analysis of the crystal structures of several homeodomain-DNA complexes has identified clear patterns between the amino acid at a particular position in the homeodomain and the DNA base pair with which it interacts. However, these patterns are only true for very related proteins, because recognition domains have distinct structures and therefore bind to the DNA at different angles.

Protein	Consensus sequence	Reference
Antennapedia	C<u>ATTA</u>	(Muller *et al.*, 1988)
bicoid	GGG<u>ATTA</u>GA	(Driever and Nusslein-Volhard, 1989)
MATα2	ACATGTAATT	(Wolberger *et al.*, 1991)
Ultrabithorax	CC<u>ATTA</u>A	(Ekker *et al.*, 1991)
AREC3/Six4	GGTGTCAGGTTGC	(Suzuki-Yagawa *et al.*, 1992)

Table 1.2 Examples of homeodomain binding sites

ATTA consensus site is underlined when present

1.7.3 Cooperative Binding of Homeodomains

Highly diverged homeodomain proteins that mediate very different transcriptional processes have been shown to bind, *in vitro*, to identical DNA sequences (Hoey and Levine, 1988). As mentioned in section 1.7.2, the majority of homeodomains bind to the core sequence ATTA. Therefore, other mechanisms are required to achieve further DNA binding specificity. Several classes of homeodomain proteins contain a second

DNA binding domain and both recognition sequences are required for specific binding.

1.7.3.1 The Paired Domain

Several homeodomain proteins, including the *Drosophila* proteins paired and gooseberry (Bopp *et al.*, 1986) and the mammalian Pax proteins which have a critical role in the development of the nervous system (Kessel and Gruss, 1990) have been shown to contain a homeodomain characterised by a serine at position 50 (most homeodomains have a lysine or a glutamine at this position) and a second 128 amino acid DNA binding domain, known as the paired domain. Both the paired domain and the characteristic homeodomain have a specific binding activity of their own and many DNA binding proteins contain only the paired domain or the homeodomain. However, when combined with a homeodomain, the paired domain increases binding specificity and the distance between the 2 DNA binding domains adds to this specificity (Treisman *et al.*, 1991).

1.7.3.2 The POU Domain

A DNA binding domain of between 67 and 70 amino acids, the POU domain is found in several homeodomain proteins and is separated from the homeodomain by a 15-27 amino acid variable spacer region (Herr *et al.*, 1988). To date the POU domain has only been found in conjunction with a homeodomain and together they recognise an 8 bp A/T rich DNA sequence, which differs between different members of the family. This octamer does not contain the ATTA core sequence. Interestingly, amino acid 50 of the homeodomain of the POU domain containing protein Pit-1, does not have an important role in determining the binding specificity of this protein (Ingraham *et al.*, 1990). High affinity binding required the presence of both binding domains. The homeodomain of mammalian Oct-1 was shown by kinetic analysis, using quantitative gel retardation assays to have a very rapid association/dissociation constant and that the POU domain increases the binding affinity of the homeodomain by increasing the half life of the retarded complex (Verrijzer *et al.*, 1992). The POU domain alone also bound to the 8 bp DNA sequence, but with low affinity and it was concluded that high affinity, sequence specific DNA binding requires both domains.

1.7.3.3 The Zinc-Finger Motif

A third method of enhancing binding specificity is likely to have been due to a recombination event between a developmental gene containing one or more homeodomains and another containing one or more zinc-finger motifs. ATBF-1 contains 4 homeodomains and 17 zinc fingers (Morinaga *et al.*, 1991). Of the 4 homeodomains 3 have a glutamine at position 50 of the homeodomain and 1 has an arginine. The homeodomain containing the arginine at position 50 is preceded by a POU domain. Together these factors greatly increase the binding specificity of the protein as many DNA sequences are required to be present in a specific order and a specific distance apart.

1.7.3.4 Dimerisation

Although homeodomain proteins will bind as monomers *in vitro*, several examples are now known where the affinity of one protein for a particular DNA sequence is greatly enhanced in the presence of the binding of a second protein. The homeodomain of the paired class (described above) and the paired-like class (that lack the characteristic serine at position 50 and are not associated with a Paired domain, but contain sequences similar to the paired homeodomain) can bind as homodimers or heterodimers to palindromic DNA sequences (Wilson *et al.*, 1993). The homeodomain of *Drosophila* paired was shown to bind as a monomer to TAATTG, but as a dimer and with stronger affinity to TAATTGATTA. The dimer is formed by the direct interaction of the 2 homeodomains which is made possible by the occurrence of a 20° bend in the DNA. Further analysis using other paired and paired-like homeodomains demonstrated the importance of the amino acid at position 50 in determining the spacing between the 2 half sites. Homeodomains with a glutamine at position 50 preferred 3 bp and those with a lysine also preferred 3 bp between the half sites but had a more stringent requirement for the identity of the base pairs. Heterodimers were also detected between paired and gooseberry which are known to be expressed at the same time in developing *Drosophila* cells.

The DNA binding specificity of the yeast homeodomain protein, MATα2, is enhanced by its association with either the homeodomain protein MATa1 or the non-homeodomain protein MCM1 (Goutte and Johnson, 1988; Goutte and Johnson, 1993; Mead *et al.*, 1996). The 2 heterodimers bind to different DNA sequences. The cooperative binding of a1/α2 relies on the 21-residue C-terminal tail of α2 binding to the a1 homeodomain. This alters the conformation of α2 which results in the DNA to

which the heterodimer binds, bending 60°. When the α2 homeodomain binds alone, it results in a negligible bend in the DNA.

The homeotic genes (*Hom*) of *Drosophila* and their mammalian homologues the *Hox* genes are essential for the correct development of diverse organisms. *In vitro Drosophila* Hom proteins Antennapedia and Ultrabithorax bind to the same 6 bp consensus sequence. However, the 2 proteins control different developmental pathways. The majority of Hom/Hox proteins have a short conserved stretch of amino acids to the N-terminal side of their homeodomains, the hexapeptide (Mann and Chan, 1996). Sequence comparison with Hox/Hom proteins between and within species suggests that the hexapeptide has evolved with the homeodomain. Members of the same species may share only 4 amino acids, whereas proteins from different species may share up to 12 amino acids in this region. *Drosophila*, extradenticle and mammalian PBX proteins (collectively known as the PBC family) interact with the hexapeptide and alter the DNA binding specificity of the homeodomain thus increasing the differences in binding specificity between members of the Hom/Hox family of homeodomain proteins.

1.7.3.5 DNA Bending Proteins

To allow transcription factors to interact with each other cofactors are often required to interact indirectly by binding to the DNA and bending it to enable other DNA bound proteins to interact with each other. The lymphoid enhancer-binding factor (LEF-1) is an example of a sequence specific DNA binding factor that can induce a sharp bend in the DNA which allows the interaction of 3 transcription factors that bind at sites flanking the LEF-1 sites (Love *et al.*, 1995).

1.7.4 Homeobox Genes and Disease

Recently, several human diseases have been shown to be linked to mutations in homeobox genes. Mammalian *paired* family Pax genes have been implicated in pathways involved in development (reviewed in Gruss and Walther, 1992) and mutations in the human homologues of the mouse *Pax* genes have been identified in a range of human diseases affecting the nervous system. Waardenburg syndrome I an autosomal dominant disorder which results in a combination of pigmentary disturbances, lateral displacement of the inner canthus of each eye, deafness and mental retardation was linked to *PAX3*. Mutations in the *PAX3* gene have now been identified in several families (Baldwin *et al.*, 1992; Tassabehji *et al.*, 1992). Pax6 and

its *Drosophila* homologue (eyeless) are known to regulate genes involved eye development (reviewed in Oliver and Gruss, 1997). Mutations in the mouse *Pax6* gene lead to a small eye phenotype and in the human homologue (*PAX6*) cause aniridia (a complete or partial absence of the iris), Peters' anomaly (malformations of the anterior chamber) or keratitis (inflammation of the cornea) (Hanson and van Heyningen, 1995; Hanson *et al.*, 1993; Mirzayans *et al.*, 1995; Ton *et al.*, 1991).

Homozygous mutations in the human POU domain encoding gene *PIT1* result in combined pituitary hormone deficiency which leads to irreversible mental and growth retardation due to a deficiency in thyroid stimulating hormone, growth hormone and prolactin (Andersen and Rosenfeld, 1994). X-linked mixed deafness has been shown to be caused by mutations in the *POU3F4* gene. Interestingly, mutations have been found between 15 and 400 kb upstream of the gene in affected individuals (De Kok *et al.*, 1995). This indicates that mutations in the regulatory elements of *POU3F4* are responsible for the disease.

Human *HOX* genes are arranged in 4 genomic clusters of approximately 100 kb in length. Each cluster contains several *HOX* genes arranged in an homologous order and genes belong to any one of 13 homology groups. These genes control the identity of the various regions along the body axis. 3' *HOX* genes are expressed early in development and the more 5' a gene is in a cluster the later it is expressed (Boncinelli, 1997). Surprisingly until very recently no severe developmental phenotypes had been linked to these important genes (Engelkamp and van Heyningen, 1996). However, several disorders are now known to be caused by mutations in *HOX* genes. In 2 unrelated families, synpolydactyly has been shown to be linked to a duplication of 9 residues in a polyalanine tract in the intron of the *HOX* gene *HOXD13* (Akarsu *et al.*, 1996). A mutation that converts a highly conserved tryptophan residue in the homeodomain of *HOXA13* to a stop codon causing the production of a truncated protein leads to hand-foot-genital syndrome (Mortlock and Innis, 1997). The number of diseases associated with homeobox genes is rapidly growing and therefore the study of these transcription factors has become increasingly important.

1.8 The *Six* Subfamily of Homeobox Genes

At the time this project was instigated 5 mammalian members and 1 *Drosophila* member of this subfamily had been identified. Recently a sixth mammalian gene (*Optx2*) and its *Drosophila* homologue (*optix*) were reported (Toy *et al.*, 1998) and a third *Drosophila* gene is currently being analysed (Graham Hamilton, personal communication). *Six* subfamily genes have now been identified in chicken, zebrafish and medaka (Bovolenta *et al.*, 1996; Bovolenta *et al.*, 1998; Loosli *et al.*, 1997; Seo *et*

al., 1998a; Seo *et al.*, 1998c; Toy *et al.*, 1998). These recent findings are discussed in section 7.3.

1.8.1 *sine oculis*

Drosophila sine oculis (*so*) was the first member of the *Six* subfamily to be identified (Cheyette *et al.*, 1994; Serikaku and O'Tousa, 1994). It is expressed in the optic lobe prior to its invagination from the embryonic ectoderm and is essential for eye development. Full length cDNAs were isolated from 4-8 hour embryonic and third instar eye disc cDNA libraries and the gene shown to encode a 2.8 kb transcript which contains an open reading frame that is predicted to encode a 416 amino acid polypeptide. *so* contains 8 exons, including an alternatively spliced untranslated first exon and the homeodomain is encoded by exons 3, 4 and 5. The homeodomain of so has diverged considerably from previously described homeodomains and shares only 68% identity over the last 2 helices (which includes the recognition helix) with its closest homeodomain outside of the Six subfamily, the homeodomain of the MAT1-P protein from the fission yeast *Schizosaccharomyces pombe* (Kelly *et al.*, 1988). Importantly, it lacks 2 highly conserved amino acids; arginine at position 5 and glutamine at position 12 of the homeodomain. Arginine at position 5 which is conserved in 97% of isolated homeodomains is known to contact base pairs in the ATTA core sequence and this indicates that so may bind to an alternative DNA sequence (Gehring *et al.*, 1994).

1.8.2 Mammalian *Six* Genes

Mouse *Six1* and *Six2* were the first of the mammalian *Six* genes to be identified and they were initially shown to be expressed during the development of the limb tendons and the head and body mesenchyme (Oliver *et al.*, 1995b). A third mammalian *Six* gene (*Six3*) was isolated from a mouse E14.5 brain cDNA library that was screened using an 800 bp fragment containing the homeobox of *Six1*. Early expression of mouse *Six3* is restricted to the anterior neural plate, with later expression in the ectoderm of the nasal cavity, the olfactory placode, Rathke's pouch, the ventral forebrain including the optic recess, the hypothalamus and the optic vesicles (Oliver *et al.*, 1995a). Mouse *AREC3 /Six4* transcripts were originally detected in skeletal muscle, heart, lung, kidney and brain (Kawakami *et al.*, 1996a). More recent developments in this field are discussed in section 7.3.

The 6 mammalian members of the Six subfamily and *Drosophila* so and optix are highly homologous in the Six domain and the homeodomain (figure 1.4). However

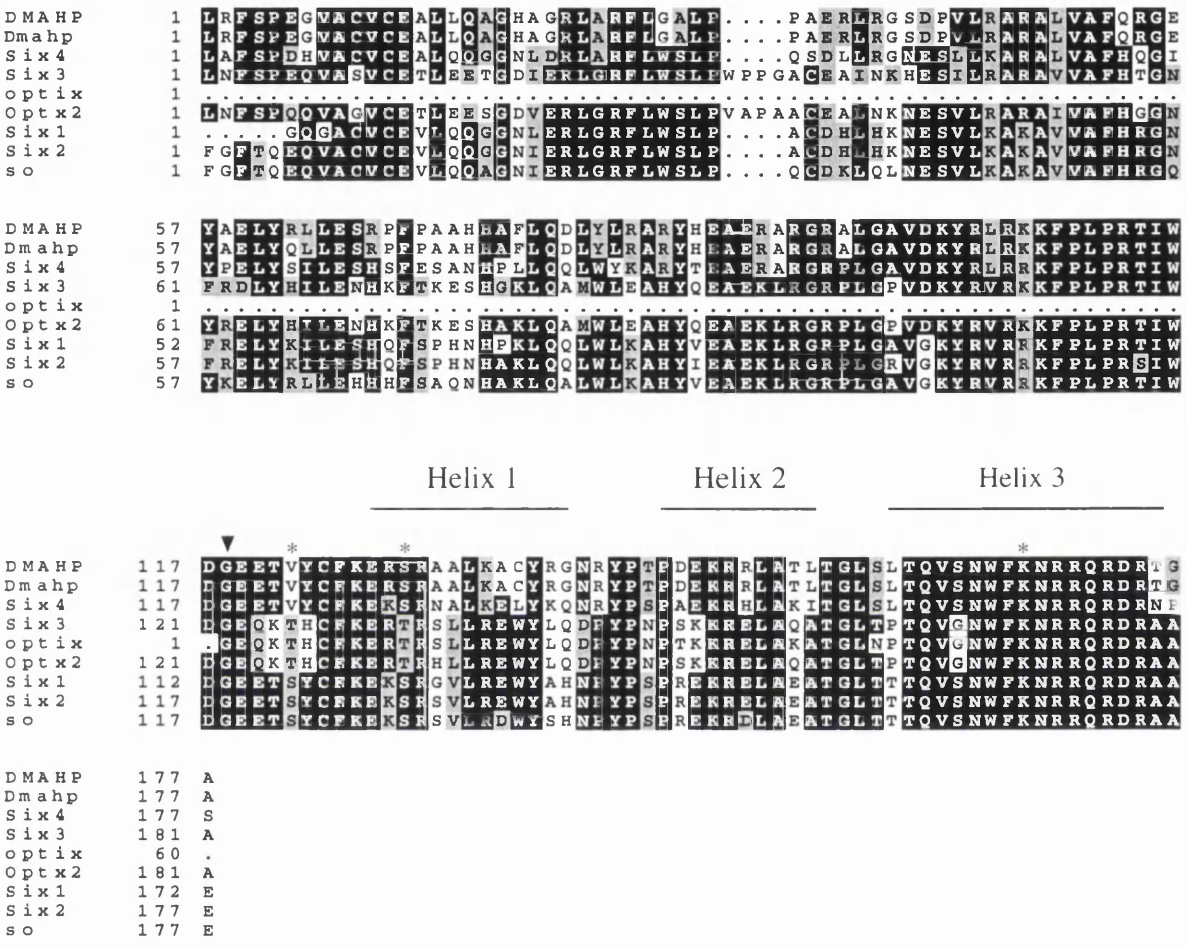

Figure 1.4 Comparison of the predicted amino acid sequences of members of the Six subfamily

The homologous regions of the predicted mouse Six subfamily proteins (Six1, Six2, Six3, Optx2, Six4/AREC3 and Six5/Dmahp), *Drosophila* so and human DMAHP have been aligned. The homeodomain of *Drosophila* optix, minus the final amino acid, is also included (no further optix cDNA sequence is available). The arrowhead indicates the start of the homeodomain and the asterisks indicate amino acids 5, 12 and 50 of the homeodomain. The residues corresponding to the 3 helices of the homeodomain are shown. The Six domain is contained within the region of homology upstream from the homeodomain. Conserved residues are highlighted in black and semi-conserved residues in grey. This figure was created using PILEUP from the GCG package on the UK Human Genome Mapping Project online computing service.

outside of these 2 domains the structures of the proteins differ greatly. The mammalian members of this subfamily also differ from known homeodomains at amino acid positions 5 and 12 of the homeodomain. The amino acid at position 5 is a valine in DMAHP/Six5 and AREC3/Six4, a serine in Six1, Six2 and so and a threonine in Six3, Optx2 and optix and the amino acid at position 12 is a serine in all members of the subfamily with the exception of Six3, Optx2 and optix in which it is a threonine (Boucher *et al.*, 1995; Cheyette *et al.*, 1994; Kawakami *et al.*, 1996a; Oliver *et al.*, 1995a; Oliver *et al.*, 1995b; Toy *et al.*, 1998). These 2 amino acids are likely to be very important with regard to the binding specificity of these proteins.

1.8.2.1 AREC3/Six4

DMAHP has a greater homology with AREC3/Six4 than with any other member of the Six subfamily. The amino acid sequence similarity is 70.3% in the Six domain and homeodomain. AREC3/Six4 was originally identified as one of the cell-type specific factors that binds to the Na^+ K^+ ATPase $\alpha 1$ subunit regulatory element (ARE) and was therefore named Na^+ K^+ ATPase $\alpha 1$ subunit regulatory element complex 3 (AREC3) (Suzuki-Yagawa *et al.*, 1992). Na^+ K^+ ATPase is the enzyme responsible for maintaining the Na^+ and K^+ gradients across the cell membrane and is composed of 2 subunits α and β. There are at least 3 isoforms of the α subunit, $\alpha 1$, $\alpha 2$ and $\alpha 3$ and the $\alpha 1$ subunit is expressed in all tissues and is important for cellular homeostasis. The ARE is a positive regulatory region situated from position -102 to -62 upstream of the transcription start site of the Na^+ K^+ ATPase $\alpha 1$ subunit gene (ATP1A1). The function of AREC3/Six4 is of particular interest as defects in Na^+ and K^+ gradients have been identified in DM patients (section 1.2).

DNase I footprinting and methylation experiments identified the binding site of AREC3/Six4 as GGTGTCAGGTTGC and a possible minimum sequence for binding as GGNGNCNGGTTGC (Suzuki-Yagawa *et al.*, 1992). A full length mouse AREC3/Six4 cDNA was obtained from a skeletal muscle cDNA library and used to construct GST-AREC3 recombinant fusion proteins containing different regions of the protein (Kawakami *et al.*, 1996a). The abilities of these deletion constructs of GST-AREC3 to bind to the ARE were tested by gel retardation assays. It was concluded by competition assays that both the Six domain and the homeodomain were required for specific binding to GGTGTCAGGTTGC although the homeodomain alone bound specifically to some other unidentified region of the ARE. This is a similar situation to that found in the Paired and POU classes of homeodomain proteins in which the presence of 2 domains is required for specific DNA binding (section 1.7.3).

1.9 Aims and General Strategy

DMAHP is a member of a subfamily of proteins that encode a homeodomain that has diverged considerably from other known homeodomains. However, many of the conserved residues that have been shown to interact with DNA are present although the conserved arginine at position 5, which when present in a homeodomain is known to contact the ATTA consensus sequence, is absent. This information along with reports that AREC3/Six4 is a cell-type specific factor that binds to the consensus sequence GGTGTCAGGTTGC within the ARE, indicates that DMAHP is a DNA and or RNA binding protein that regulates the expression of other genes.

It was decided to investigate the DNA binding properties of recombinant DMAHP fusion proteins and to initiate the identification of genes regulated by DMAHP. Therefore, the aims of the project were:

- To isolate a full length human *DMAHP* cDNA.
- To use a bacterial expression system to overexpress and affinity purify GST-fusion proteins containing the Six domain, the homeodomain and both domains together.
- To investigate the binding activities of these 3 fusion proteins by gel retardation assays using double stranded oligonucleotides representing putative DNA binding sites. (It was hypothesised that the Six domain would have DNA binding properties distinct from the homeodomain and that both domains would be required for specific DNA binding. Putative DNA binding sites were predicted from the known binding sites of AREC3/Six4 and classical homeodomain proteins).
- To develop and use a whole genome PCR based screen to identify DNA sequences to which DMAHP binds, as an initial stage in the identification of genes which it regulates.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Bacterial Strains

The *Escherichia coli* bacterial strains used, are listed in table 2.1 along with their genotype and supplier.

Strain	Genotype	Supplier
DH5α	F-, φ80d*lac* ZΔM15, Δ(*lac* ZYA-*arg*F), U169, *deo*R, *rec*A1, *end*A1, *hsd*R17(r_K-,m_{K+}), *sup*E44, λ^-, *thi*-1, *gyr*A96, *rel*A1.	Life Technologies Ltd.
Jm109	*rec*A1, *end*A1, *gyr*A96, *thi*, *hsd*R17(r_K-,m_{K+}), *rel*A1, *sup*E44, Δ(*lac-pro*AB), [F', *tra*D36, *pro*AB, *lac*IqZΔM15]	Promega
MOS*Blue*	*end*A1, *hsd*R17(r_K-,m_{K+}), *sup*E44,*thi*-1, *gyr*A96, *rel*A1, *lac*, [F'*pro*AB, *lac*IqZΔM15, Tn10(tet^r)], *rec*A1	Amersham International plc.
TOP10	*mrc*A, Δ (*mrr hsd*RMS*mcr BC)*, F80D*lac* ZΔM15, Δ*lac* X74, *deo*R, *rec*A1, *ara* D139, Δ (*ara, leu*), 7697, *gal* U, *gal*K, I-, *rps* L, *end*A1, *nup*G, F^-	Invitrogen
XL-Blue	*sup*E44, *hsd*R17, *thi*-1, *gyr*A96, *end*A1, *rec*A1, *rel*A1 *lac*- F'*pro*AB^+, *lac*IqZΔM15 Tn*10* (Tet^r)	Stratagene

Table 2.1 Bacterial Strains

2.1.2 Chemicals and Reagents

AnalaR and molecular biology grade chemicals were obtained from Merck Ltd. (BDH Laboratory Supplies) and Sigma Chemical Company Ltd, with the exception of the items in table 2.2.

Supplier	Chemicals and reagents
BIO-RAD	Kaleidoscope prestained standards, 30% (w/v) acrylamide/bis 29:1 (3.3% C) and 30% (w/v) acrylamide/bis 37.5:1 (2.6% C)
Boehringer Mannheim	Fraction V, fatty acid free, BSA, agarose, IPTG, X-gal, PMSF, trizma base
Calbiochemica®	Nonidet® P-40
Difco Laboratories	Agar, Bacto® tryptone, Bacto® yeast extract (without amino acids)
Eastman Kodak Company	Formamide
Fisher Scientific U.K. Ltd	Acetic acid, boric acid, chloroform
Fluka Biochemica	Guanidine hydrochloride
FMC BioProducts	NuSieve agarose
Genetic Research Instrumentation	Developer and fixer
Kodak Scientific Imaging Systems	LMP agarose
ICN	[α-^{32}P] dCTP (3000 Ci/nmol) and [γ-^{32}P] ATP (3000 Ci/nmol)
Life Technologies Ltd	1 kb DNA ladder
Pharmacia Biotech	Dextran sulphate, dNTPs, Ficoll®400, Glutathione, Glutathione Sepharose®4B, Phastgel™ Blue R, random hexamers, Sephadex® G-50 and Sephadex® G-25, poly[dI/dC].poly[dI/dC]
Premier Beverages	Skimmed milk powder
Severn Biotech Ltd	Acrylamide premix

Table 2.2 Chemicals and Reagents

in Ltd. Inoculating loops and 90 mm Petri
:ientific and 15 ml and 50 ml Falcon Tubes

e (pMOS-DMAHP-ABC) and cosmid
ɛr (Keith Johnson's group).

:s were obtained from Boehringer
ɛw England Biolabs Ltd. Bacteriophage T4
IAP were obtained from Promega. The
vas obtained from Boehringer Mannheim.
ɔtained from Sigma Chemical Company Ltd.
ringer Mannheim, Promega and Perkin

2.1.6 Immunochemicals

Polyclonal goat anti-*Schistosoma japonicum* GST serum was obtained from Pharmacia Biotech and polyclonal donkey anti-goat IgGs conjugated with HRP from Serotec. Polyclonal rabbit anti-DMAHP 2 R254„ GST depleted serum was a gift from Catherine Winchester (Keith Johnson's group). Polyclonal donkey anti-rabbit IgGs conjugated with HRP were obtained free of charge from SAPU.

2.1.7 Kits

The following kits were used during the project (table 2.3).

Supplier	Kit
Amersham International plc.	ECL™ western blotting analysis system, pMOS*Blue* T-vector kit
Millipore	Centricon 10 membrane units
Perkin Elmer	ABI Prism™ Dye Terminator Cycle Sequencing Ready Reaction Kit
Promega	pGEM® -T-Easy vector system
Scotlab	Nucleon II genomic DNA preparation kit
Qiagen	Plasmid DNA mini, midi and maxi kits, gel extraction DNA purification kit

Table 2.3 Kits

2.1.8 Membranes and Paper

Hybond™-N nylon membrane and Hybond™-N ECL nitrocellulose membranes were obtained from Amersham International plc. 3MM chromatography paper was supplied by Whatman and blot absorbent filter paper by BIO-RAD. Saran wrap was obtained from Dow.

2.1.9 Oligonucleotides

Oligonucleotides for both DNA sequencing and for use as DNA fragments in gel retardation assays were supplied by Genosys, Perkin Elmer and Cruachem. Sequences are given in table 2.4.

Oligonucleotide	Sequence 5'-3'
T7	TAATACGACTCACTATAGGG
T3	AATTAACCCTCACTAAAGGG
SP6	GATTTAGGTGACACTATAG
pGEXF1	GGCGACCATCCTCCAAAATC
pGEXB1	TTTTCACCGTCATCACCGAAAC
U-19mer	GTTTTCCCAGTCACGACGT
DMAHPFB	ATGTGCACCTCATCAACTCC
DMAHPRC	TTAGTTCCAGCCCTGCTGAC
CBR005	AGGACCTGGGAGACGAGCAT
CBR0013	AGGAGGCTCAGTCGGAGG
KJDME2R	GCAGGGCAAGATCATCCTCA
DMAHPF	AGTGGACAAGTATCGACT
DMAHPRSEH	GTCCAGCCGAAGTCCTGAGG
V-SU	GATCGGACTTGCTACGGTAATCAG
TSP	AATTGGACTTGCTACGGTAATCAG
V-NL	CTGATTACCGTAGCAAGTCC
ARE1	GATCCCCGGTGTCAGGTTGCTCCG GTAACGGTGACGTGCG
ARE2	AATTCGCACGTCACCGTTACCGGAG CAACCTGACACCGGG
AREmut1	GATCCCCGGTGTGAGGTTGCTCCG GTAACGGTGACGTGCG
AREmut2	AATTCGCACGTCACCGTTACCGGAG CAACCTCACACCGGG
ARE-like1	GATCCTGGAAGCCCCTCAGGTTCCT CCCTGTCCCGCAGGG
ARE-like2	AATTCCCTGCGGGACAGGGAGGAA CCTGAGGGGCTTCAAG
GGATTA1	GATCCTGTTTCTGTTTCTAATCCCAG CCTGGGAAGGAATG
GGATTA2	AATTCATTCCTGCCCAGGCTGGGAT TAGAAACAGAAACAG

Table 2.4 Oligonucleotides

2.1.10 Photography and Autoradiography

Agarose gels were photographed using UVP gel documentation system 7500. Fugi RX X-ray film was obtained from Genetic Research Instrumentation Ltd and developed using an X-Ograph Ltd. X-Ograph Compact X2.

2.1.11 Plasmid cDNA Libraries

The pCDM8 human foetal brain and foetal muscle cDNA libraries were a gift from David Simmons (Cell Adhesion Laboratory, ICRF, IMM, John Radcliffe Hospital, Oxford, OX3 9DU). The pcDNAII human foetal brain cDNA library was a gift from Kay Davies (Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, OX3 9DU).

2.1.12 Vectors

The vectors used for subcloning regions of DNA are listed in table 2.5.

Supplier	Vector
Amersham International plc	pMOS*Blue* T-vector
Pharmacia Biotech	pGEX4T3 bacterial expression vector
Promega	pGEM®-T Easy vector
Stratagene	pBluescript® SK(+) phagemid

Table 2.5 Vectors

2.2 Solutions and Media

Solutions and media were sterilised by autoclaving at 121°C for 15 minutes or by filtration through a sterile 0.2 μm pore filter obtained from Gelman Science.

2.2.1 Solutions

Denaturing solution

1.5 M NaCl, 0.5 M NaOH.

100 x Denhardt's solution (Denhardt, 1966)

2% (w/v) Ficoll®400, 2% (w/v) polyvinylpyrolidine, 2% (w/v) BSA.

5 x DNA loading dye

0.5% (w/v) SDS, 0.25% (w/v) xylene cyanol, 0.25% (w/v) bromophenol blue, 1.5% (w/v) Ficoll®400, 3 x TBE.

DL-Dithiothreitol (DTT)

Stock solution: 0.1 M in H_2O (stored at -20°C).

Working solution: 1-5 mM.

DNase I dilution buffer

50 mM Tris-HCl (pH 8.0), 50 mM NaCl, 0.1 mM EDTA (pH 8.0), 50% (v/v) glycerol

DNase I stop solution

95% (v/v) formamide, 20 mM EDTA (pH 8.0), 0.05% (w/v) bromophenol blue. 0.05% (w/v) xylene cyanol.

1.5 x DNA-protein binding buffer

50 mM Tris-HCl (pH 8.0), 120 mM NaCl, 10 mM DTT, 1 mM PMSF.

0.5 M EDTA pH 8.0

0.5 M EDTA with NaOH to pH 8.0.

Ethidium Bromide

Stock solution: 10 mg/ml in H_20.

Working solution: 200 ng/ml.

Hybridisation solution (for Southern analysis)

5 x Denhardt's solution, 5 x SSPE, 0.1% (w/v) SDS, 6% (w/v) PEG, 200 µg/ml sheared and denatured salmon sperm DNA.

1 kb ladder

60 ng/µl 1 kb ladder, 1 x DNA loading dye, 1 x TBE.

10 x Lysogen Extraction Buffer

0.5 M Tris-HCl (pH 7.5), 5 mM EDTA (pH 8.0).

Diluted to 1 x strength and 1 mM-5 mM DTT and 1 mM PMSF added, prior to use.

Manual Sequencing Gel Mix

8.3 M Urea, 6% (w/v) acrylamide/bis premix (19:1), 1 x TBE, 0.05% (w/v) APS, 0.05% (v/v) TEMED.

MTPBS Buffer

5 x PBS, 0.5 M NaCl, 0.05 M EDTA (pH 8.0), 5 mM PMSF.

Neutralising solution

1.5 M NaCl, 0.5 M Tris-HCl (pH 6.5).

Non-denaturing Polyacrylamide Gel Mix (Low-ionic strength)

0.007 M Tris-HCl (pH 7.9), 0.002 M EDTA (pH 8.0), 8% (w/v) acrylamide/bis (37.5:1), 2.5% (v/v) glycerol, 0.075% (w/v) APS, 0.085% (v/v) TEMED.

Non-denaturing PAGE (Low-ionic strength) Running Buffer

0.007 M Tris-HCl (pH 7.9), 0.003 M sodium acetate, 0.001 M EDTA (pH 8.0).

Non-denaturing Polyacrylamide Gel Mix (1 x Tris/Glycine)

0.05 M Tris.Glycine (pH 9.4), 0.0001 M EDTA (pH 8.0), 8% or 4% (w/v) acrylamide/bis (37.5:1), 0.12% (w/v) APS, 0.06% (v/v) TEMED.

Non-denaturing PAGE (1 x Tris/Glycine) Running Buffer

0.05 M Tris.Glycine (pH 9.4), 0.0001 M EDTA (pH 8.0).

Oligo-labelling buffer (OLB)

30% (v/v) TE pH 8.0, 30 OD units/ml of random hexamers, 1 M HEPES (pH 6.6), 0.35% (v/v) β-mercaptoethanol, 0.1 mM dATP, 0.1 mM dGTP, 0.1 mM dTTP, 0.24 M Tris-HCl (pH 8.0), 0.0024 M $MgCl_2$.

Orange G loading dye

0.06% (w/v) Orange G, 50% (v/v) glycerol

Phenylmethylsulphonyl Fluoride (PMSF)
Stock solution: 100 mM in propan-2-ol (stored at -20°C).
Working solution: 0.1-1 mM

10 x Phosphate Buffered Saline (PBS)
1.4 M NaCl 0.027 M KCl, 0.1 M Na_2HPO_4, 0.021 M KH_2PO_4 (pH 7.2 with HCl).

PBS-T
1 x PBS, 0.1% (v/v) Tween®20.

10% (w/v) SDS
10% (w/v) SDS in H_2O.

SDS-Polyacrylamide Separating Gel mix
0.375 M Tris-HCl (pH 8.7), 0.1% (w/v) SDS, 10% or 15% (w/v) acrylamide/bis (29:1), 0.05% (w/v) APS, 0.05% (v/v) TEMED.

SDS-Polyacrylamide Stacking Gel mix
0.125 M Tris-HCl (pH 6.8), 0.1% (w/v) SDS, 4% (w/v) acrylamide/bis (29:1), 0.05% (w/v) APS, 0.1% (v/v) TEMED.

2 x SDS-PAGE Loading Buffer
0.125 M Tris-HCl (pH 6.8), 20% (v/v) glycerol, 2% (w/v) SDS, 0.5% (v/v) β-mercaptoethanol, 0.025% (w/v) bromophenol blue.

5 x SDS-PAGE Loading Buffer
0.313 M Tris-HCl (pH 6.8), 50% (v/v) glycerol, 5% (w/v) SDS, 1.25% (v/v) β-mercaptoethanol, 0.0625% (w/v) bromophenol blue.

5 x SDS-PAGE Running Buffer (Mini Protean II Apparatus)
0.25 M Tris, 0.5% (w/v) SDS, 1.85 M glycine.

20 x SSC
3 M NaCl, 0.3 M $Na_3C_6H_5O_7.2H_2O$.

20 x SSPE
0.14 M $Na_2HPO_4.2H_2O$, 0.06 M $NaH_2PO_4.H_2O$, 3.6 M NaCl, 0.01 M EDTA (pH 8.0).

1 x TBE

0.009 M Trizma base, 0.09 M orthoboric acid, 2 mM EDTA (pH 8.0).

TE buffer (TE)

10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0).

1 M Tris pH 8.0

1 M Trizma base and HCl to pH 8.0.

1 M Tris pH 7.9

1 M Trizma base and HCl to pH 7.9

1 M Tris pH 7.5

1 M Trizma base and HCl to pH 7.5

1 M Tris pH 6.8

1 M Trizma base and HCl to pH 6.8

1.5 M Tris pH 8.7

1.5 M Trizma base and HCl to pH 8.7

2.2.2 Bacterial Solutions, Media and Antibiotics

Ampicillin

Stock solution of 50 mg/ml in H_2O, stored at -20°C.

Working solution of 50 µg/ml.

IPTG

Stock solution of 100 mg/ml in H_20, stored at -20°C.

Working stock of 10 µg/ml.

Luria-Bertani (LB) medium

1% (w/v) Bacto® tryptone, 0.5% (w/v) Bacto® yeast extract, 1% (w/v) NaCl.

LB agar contained LB with 1.5% (w/v) agar.

SOB medium

2% (w/v) Bacto® tryptone, 0.5% (w/v) Bacto® yeast extract, 0.85 mM NaCl, 0.25 mM KCl (pH 7.0 with NaOH). Sterilised 10 mM $MgSO_4$ added prior to use.

Tetracycline

Stock solution of 100 mg/ml in 50% (v/v) ethanol, stored at -20°C.
Working concentration of 10μg/ml.

X-gal

Stock solution of 50 mg/ml in dimethylformamide, stored at -20°C.
Working solution of 50 μg/ml.

2YT medium

1.6% (w/v) Bacto® tryptone, 1% (w/v) Bacto® yeast extract, 1% (w/v) NaCl.

2.3 Methods

2.3.1 Preparation of DNA

2.3.1.1 Preparation of Plasmid and Cosmid DNA

Minipreparations of plasmid DNA (for use in diagnostic tests of insert size) were performed using an alkaline lysis method (Birnboim and Doly, 1979; Ish-Horowicz and Burke, 1981). 5 ml LB or 2YT, 50μg/ml Ampicillin, 1μl glycerol stock or 1 single colony was incubated for 18 hours, shaking at 37°C. 1.5 ml of culture was centrifuged at 10,000 g for 10 minutes and the supernatant discarded. The pellet was resuspended in 100μl of 50 mM glucose, 10 mM EDTA, 25 mM Tris (pH 8.0). 200 μl of fresh 1% SDS and 200 mM of NaOH was added, mixed gently and incubated on ice for 5 minutes. 150 μl of 3 M CH_3COOK (pH 5.5) was added, mixed immediately but gently and incubated in ice for 5 minutes. The sample was centrifuged at 10,000 g for 15 minutes and the supernatant removed to a new tube. A phenol/chloroform extraction was then performed. 1 ml of 100% (v/v) ethanol was added and incubated at -70°C for 10 minutes before being centrifuged at 10,000 g for 15 minutes. The supernatant was removed and the pellet washed in 500 μl of 70% (v/v) ethanol before being air dried and resuspended in 20 μl H_20 and 0.025 μg/μl RNaseA.

Qiagen kits were also used, according to manufacturer's instructions, for the preparation of plasmid and cosmid DNA and were used at all times when the purity of the DNA was essential.

2.3.1.2 Preparation of Genomic DNA

DNA was extracted from human blood using Nucleon II kits, according to manufacturer's instructions .

2.3.2 Agarose Gel Electrophoresis

DNA molecules were separated according to size by agarose gel electrophoresis. Solutions of 1.0-3.5% (w/v) agarose, Nusieve agarose or LMP agarose in 1 x TBE were prepared in a microwave. The solution was allowed to cool to 50°C and then 200 ng/ml ethidium bromide was added and the gel cast in a horizontal tray with a 1.5 mm comb. DNA samples were mixed with 5 x DNA or Orange G loading dye and electrophoresed in 1 x TBE for 1 hour to overnight at 2-10 V/cm. 600 ng of a 1 kb ladder was also run as a DNA size marker. DNA samples were visualised using a UV transilluminator (wavelength 254 nm) and photographed.

2.3.3 DNA Purification

2.3.3.1 Phenol/Chloroform Extraction of DNA

Protein and salt contaminants were removed from plasmid DNA by the addition of an equal volume of 1:1 (v/v) phenol/chloroform. The mixture was vortexed and then centrifuged at 10,000g for 5 minutes. The aqueous phase, containing the DNA, was transferred to a fresh tube and an equal volume of chloroform added, to remove any remaining traces of phenol. The mixture was vortexed and then centrifuged at 10,000g for 2 minutes. The aqueous layer was then transferred to a clean tube and the DNA precipitated with ethanol.

2.3.3.2 Ethanol Precipitation of DNA

2.5 volumes of 100% (v/v) ethanol and 0.1 volume of 3 M CH_3COONa (pH 5.2) were added to a DNA solution. The mixture was vortexed and placed at -70°C or in ice for 10 minutes, before being centrifuged for 15 minutes. The pellet was washed in 70% (v/v) ethanol, to remove excess salt and then air dried and resuspended in a suitable volume of 1 mM Tris (pH 8.0) or H_2O.

2.3.3.3 Purification of DNA from Agarose Gels

2.3.3.3.1 Hot Phenol Extraction

DNA fragments were cut out of a low melting point agarose or Nusieve agarose gel and weighed. Assuming a density of 1 g/ml, 3 volumes of H_2O were added and the gel melted a 65°C. An equal volume of phenol was added, mixed, incubated for 3 minutes at 65°C and centrifuged at 10,000g for 5 minutes. The top aqueous layer was phenol/chloroform extracted and the DNA ethanol precipitated.

2.3.3.3.2 Qiagen Gel Extraction

Qiagen gel extraction kits were also used, according to manufacturer's instructions, to purify DNA from agarose gels.

2.3.4 DNA Concentration

DNA samples were diluted 5 µl in 750 µl H_2O. A spectrophotometer was baseline corrected using H_2O before the sample was analysed in the range 200 to 300 nm. DNA concentration was calculated as [50 x 750 x OD (at 260 nm)]/5 µg/ml.

2.3.5 DNA Sequencing

DNA sequencing was performed on either an ABI 373A sequencer or an ABI 377 sequencer, by the Molecular Biology Support Unit at the University of Glasgow. Reactions were performed as stated in the ABI PRISM™ Dye Terminator Cycle Sequencing Ready Reaction Kit protocol. Ethanol precipitation protocol 1 was used for purifying extension products.

2.3.5.1 Computer Analysis of DNA Sequences

DNA sequences were edited and aligned using Sequence Navigator, version 1.0 (Applied Biosystems). MacVector, version 4.1.4 (Kodak Scientific Imaging Systems) was used to predict the molecular weights of the recombinant fusion proteins. Database searches of GenBank, EMBL, DDBJ and PDB were performed at the National Centre for Biotechnology Information using BLASTN network services, accessed via the UK Human Genome Mapping Project online computing service.

DNA sequences were analysed for restriction sites using MAP, and compared to find sequence homology using BESTFIT and LALIGN from the GCG package on the UK Human Genome Mapping Project online computing service.

2.3.6 Polymerase Chain Reaction (PCR)

Except where stated PCRs contained 0.25 μM each primer, 0.2 mM each dNTP 1.25U *Taq* polymerase 1-5% (v/v) DMSO and 1x buffer in a 25 μl total volume. Reactions were thermal cycled under the following conditions on either a Perkin Elmer DNA Thermal Cycler 480 or Gene Amp® PCR System 9700:

1 cycle:	94°C for 4 minutes
30 cycles:	94°C for 1 minute
	55-60°C for 1 minute
	72°C for 2 minutes
1 cycle:	72°C for 7 minutes

2.3.7 Southern Blotting

Gels were incubated in denaturing solution for 45 minutes, neutralising solution for 45 minutes, and the capillary blot was assembled using Hybond™-N nylon membrane. All sheets of Hybond™-N and 3 MM paper were rinsed in 20 x SSPE before assembly, and care was taken to eliminate any air bubbles. Gels were blotted for 16 hours, and then the membrane was baked at 80°C for 2 hours.

2.3.8 Preparation of Radiolabelled Double Stranded probes

20-50 ng DNA and H_2O to a final volume of 32 μl was denatured at 100°C for 10 minutes, and briefly quenched on ice. 10 μl OLB, 2μl 1 mg/ml BSA, 5 μl [α-^{32}P] dCTP (3000 Ci/mmole) and 2U Klenow were added, and incubated at room temperature for 3 hours.

A 1 ml Sephadex® G-50 column was prepared in a 1.5 ml eppendorf with polymer wool and 4% (w/v) Sephadex® G-50 in TE. The sephadex was compacted by centrifuging twice at 1,800 rpm for 4 minutes. 50 μl 30 mM NaCl in TE was added to the column to elute the radiolabelled probe, and centrifuged at 1,800 rpm for 4 minutes. The eluate was collected and the incorporation as a percentage of the total was calculated by taking cpm readings before and after purification.

2.3.9 Southern Hybridisation (using double stranded probes)

Filters to be hybridised were place in a hybridisation bottle with 20 ml of hybridisation solution and rotated at 65°C for 3 hours. The purified radiolabelled probe was denatured at 100°C for 10 minutes and quenched briefly on ice before being added to the bottle. Hybridisation was carried out at 65°C, overnight in a rotating hybridisation oven.

Filters were washed, in 0.1 % SDS, sequentially to the following stringency's:

2 x SSPE	room temperature	20 minutes
2 x SSPE	65°C	20 minutes
1 x SSPE	65°C	20 minutes
0.1 x SSPE	65°C	20 minutes

At each stage the filter was monitored with a Geiger counter, and washing was stopped if the reading was below 10 cps. Filters were wrapped in Saran wrap whilst still damp, and exposed to X-ray film in the presence of an intensifying screen at -70°C. Autoradiographs were developed after overnight - 3 day exposure.

2.3.10 DNA Cloning Techniques

2.3.10.1 Restriction Enzyme Digests of Plasmid DNA

1 µg DNA was incubated with 5-10U restriction endonuclease in 1 x restriction endonuclease buffer (as recommended by the manufacturer) in a final volume of 20 µl at the appropriate temperature for 4 hours.

2.3.10.2 Calf Intestinal Phosphatase Treatment (CIAP) - to remove 5' phosphate groups from DNA

1 µg digested DNA was CIAP treated in 80 µl volume with 2 µl 1:10 dilution CIAP enzyme (5U) in 1 x manufacturer's supplied buffer at 37 °C for 1 hour.

2.3.10.3 Klenow Treatment - to fill recessed 3' termini

1U of the Klenow fragment of DNA polymerase I and 1 µl 1 mM each required dNTP were added to a 20 µl digest and incubated at room temperature or 37°C for 15

minutes. The reaction was terminated by heating to 75°C for 10 minutes or by phenol/chloroform extraction.

2.3.10.4 Ligations

A total of 25-30 ng DNA was ligated with 1U DNA ligase in 1 x ligase buffer (supplied by the manufacturer) in a 10 µl volume at 4°C overnight. A molar insert to vector ratio of 3:1 was used with sticky ended ligations and of 5:1 for blunt ended ligations.

2.3.10.5 Preparation of Competent Cells

A 50 ml culture inoculated with 500 µl overnight culture was grown with aeration in medium A [LB with 10 mM $MgSO_4.7H_20$, 0.2% (w/v) glucose] to mid logarithmic phase. The cells were kept on ice for 10 minutes, then pelleted at 1500g for 10 minutes at 4°C. The cells were resuspended gently in 500 µl of pre-cooled medium A, then 2.5 ml storage solution B [36% (w/v) glycerin, 12% (w/v) PEG (MW7500), 12 mM $MgSO_4.7H_20$ added to LB (pH 7.0)] was added and mixed well without vortexing. The competent cells were divided into aliquots of 100 µl each and stored at -80°C until use.

2.3.10.6 Transformations

100 µl competent cells for each transformation were thawed for 10 minutes on ice. 5 µl ligation mixture was added and the tube tapped gently to mix. They were then incubated on ice for 30 minutes. The cells were then heat shocked at 42°C for 90 seconds and cooled on ice for 2 minutes. 500 µl pre-warmed LB was added and the samples incubated at 37°C for 1 hour. 30 µl of each transformation was spread on an LB plate containing the appropriate antibiotics and where appropriate 30 µg/ml IPTG and 50 µg/ml X-gal and allowed to dry for 10 minutes, before being incubated overnight at 37°C.

2.3.10.7 Generation of Plasmid Stocks

Bacteria transformed with recombinant plasmids were prepared for long term storage by inoculating 5ml of 2YT or LB medium containing 50 µg/ml ampicillin with a single colony from an agar plate. This was incubated overnight, shaking at 37°C. 500 µl of

overnight culture was mixed with 500 µl of 40% (v/v) glycerol, 60% (v/v) 2YT medium and stored at -70°C.

2.3.11 Small Pool PCR-Based Screening of cDNA Libraries

2.3.11.1 Pooling cDNA Libraries

XL-1 BLUE competent cells were transformed (section 2.3.10.6) with 1 x 10^6 clones from a plasmid cDNA library. Following 1 hour incubation at 37°C, 6.4 ml of pre-warmed LB and 50 µg/ml ampicillin were added to the transformation. The transformed cells were divided equally between 64 wells in an 8 x 8 matrix, in a microtitre plate (100 µl of culture, containing 15,625 individual, unamplified clones per well) and allowed to grow overnight at 37°C, shaking at 150 rpm. The microtitre plate was placed in a sandwich box containing damp paper towels to prevent the cultures drying out. The wells from each row and column were pooled to give 16 individual pools, each pool containing 25 µl of its constituent wells. An equal volume of 60% 2YT/40% glycerol (v/v) was added to each pool and well and they were stored at -70°C.

1 µl glycerol stock of each pool was used, to prepare 5 ml overnight cultures. Plasmid DNA from 1.5 ml of each culture was prepared (section 2.3.1.1) and resuspended in 30 µl of H_20, before being diluted 1/100 in H_20. 1 µl of diluted DNA from each pool was used as a template in a PCR. 70% (v/v) of each PCR was electrophoresed in a 2% (w/v) agarose gel. DNA was transferred to Hybond™-N membrane and hybridised to a radiolabelled DNA probe (sections 2.3.7-2.3.9). DNA from positive wells was prepared and 1 µl of a 1/100 dilution used as a template in another round of PCR. 70% (v/v) of each PCR was electrophoresed in a 2% (w/v) agarose gel and the gel was Southern blotted and probed with a radiolabelled DNA probe.

To determine the number of clones in a positive well 1µl of glycerol stock was diluted in LB to give dilutions of 1 x 10^{-4}, 1 x 10^{-5} and 1 x 10^{-6} and spread on LB plates containing 50 µg/ml ampicillin and incubated overnight at 37°C. 40,000 clones from a positive well were divided between 64 wells (100 µl of culture, containing 625 clones per well) in a microtitre plate. The clones were amplified and pooled as described above and analysed by PCR. Clones from a positive well were screened by hybridisation to a radiolabelled probe (sections 2.3.11.2 -2.3.11.4).

2.3.11.2 Plating Out Plasmid Library

Hybond™ -N+ was placed on the surface of 6 LB agar + ampicillin plates. 500 clones were plated out on the surface of each membrane and incubated at 37°C for 24 hours. Replica plates were made by laying a master filter colony side down onto a second Hybond™-N+ filter, placed on a sheet of 3MM paper and then covering with a second sheet of 3MM paper. They were then pressed between two glass plates. Orientation marks were made through the Hybond N™+ filters using a sterile needle. Master filters were then incubated on fresh plates for 6 hours and replica filters for 10 hours. After 6 hours a second set of replica filters was made and both master and replica plates were incubated for 4 hours after which the second set of replica plates were left at room temperature overnight. The master plates and first set of replica plates were stored at 4°C overnight. The replica filters were then placed colony side up on 3MM paper wetted with denaturing solution and left for 7 minutes and then neutralised in the same way for 7 minutes. The filters were then immersed in 2 x SSC and cell debris was removed by rubbing the filter gently before being rinsed in fresh 2 x SSC and baked at 80°C.

2.3.11.3 Preparation of Radiolabelled Oligonucleotide Probes

10 pmoles of oligonucleotide were mixed with 10U polynucleotide kinase, 10 μCi [γ-^{32}P] ATP and 1x kinase buffer in a 20μl final volume and incubated a 37°C for 2 hours.

A 1 ml Sephadex® G-25 column was prepared in a 1.5 ml eppendorf with polymer wool and 4% (w/v) Sephadex® G-25 in TE. The sephadex was compacted by centrifuging once at 1,800 rpm for 5 minutes. The radiolabelled oligonucleotide was added to the column and centrifuged at 1,800 rpm for 4 minutes. The eluate was collected and the percentage incorporation was calculated by taking cpm readings before and after purification.

2.3.11.4 Hybridisation of Replica Filters

The filters were pre-hybridised for 3 hours in a sealed petri dish at 53°C. An oligonucleotide probe was radiolabelled and added to 12 ml prewarmed hybridisation mix. Filters were transferred to a new petri dish with 1 ml of probe and hybridisation mix inbetween each layer. The petri dish was sealed and hybridisation was carried out at 53°C overnight.

Filters were washed twice in 6 x SSPE, 0.1% SDS for 5 minutes at room temperature. Filters were wrapped in Saran wrap whilst still damp, and exposed to X-ray film in the presence of an intensifying screen at -70°C.

2.3.12 Protein Expression

2.3.12.1 Generation of Recombinant Proteins

Recombinant protein was generated in a bacterial expression system for use in gel retardation assays and a whole genome PCR based screen, to determine the target DNA binding sites of the recombinant protein. A pGEX bacterial expression vector, that allows inducible expression by IPTG under the control of the *tac* promoter, was used. The gene fragments to be expressed were subcloned in frame into the MCS, which is situated downstream of *Schistosoma japonicum* Glutathione S-transferase (GST). GST fusion proteins were purified by affinity chromatography with glutathione as the ligand. Diagnostic restriction digests and sequencing were used to check for the correct insert and reading frame.

2.3.12.2 Optimisation of Protein Expression

2 ml SOB medium containing 50 μg/ml ampicillin was inoculated with a single colony of appropriate recombinant bacteria and incubated overnight at 37 °C. A culture was also grown that contained a vector only clone, as a positive control. 100 ml SOB medium containing 50 μg/ml ampicillin was then inoculated with 400 μl overnight culture and incubated at 37°C until an OD_{600} of 0.6 was reached. 1 ml of culture was then centrifuged at 10,000g for 10 minutes and the supernatant discarded. The pellet was resuspended in 100 μl 1 x PBS and stored at -20°C. The remainder of the culture was split into 25 ml fractions and induced with various concentrations of IPTG (0-5 mM). The cultures were then incubated at either 30°C or 37°C for up to 30 hours. The OD_{600} was monitored at hourly intervals with 1 ml of culture. 1 ml of each culture was also removed to determine the amount of protein expression by SDS-PAGE. It was centrifuged at 10,000g for 10 minutes and the supernatant removed. The pellet was resuspended in 100 μl 1x PBS and stored at -20°C. Incubation was continued overnight and the culture was sampled once again. The cells from each time point were lysed by freezing in a dry ice/methanol slurry and incubating briefly, at 42°C to rapidly thaw. This was repeated 6 times and the samples were centrifuged at 10,000g for 10 minutes. The supernatant was removed and both supernatant and pellet were snap frozen on dry ice. 100 μl of 2 x SDS loading buffer was added to all samples

and then they were boiled at 100°C for 10 minutes to denature the proteins before SDS-PAGE analysis. Where no overexpressed protein was visible by Coomassie Blue staining it was detected by western blot analysis.

2.3.12.3 Large Scale Production of Recombinant GST Fusion Proteins

5 ml SOB medium containing 50 μg/ml ampicillin and 2% (w/v) glucose was inoculated with a single colony of appropriate recombinant bacteria and incubated overnight at 37°C. 500 ml SOB medium containing 50 μg/ml ampicillin and 2% (w/v) glucose was then inoculated with 400 μl overnight culture and incubated at 37°C until and OD_{600} of 0.6 was reached. 1 mM IPTG was added to the culture to induce protein expression and incubation was continued at 30°C for the pre-determined optimum time. The cells were harvested in 200 ml fractions at 10,000g for 10 minutes at 4°C and the supernatant discarded. The pellet was resuspended in 10 ml of 1 x lysogen extraction buffer containing 5 mM DTT and 0.1 mg/ml lysozyme. The cells were lysed by sonicating on ice 5 times, each time for 10 seconds, with a 1 minute incubation on ice between bursts. At this stage, the samples were either pooled and centrifuged at 7740g at 4°C for 10 minutes and the supernatant snap frozen and stored at -70°C or 1% (v/v) Triton X-100 was added and incubated at room temperature for 30 minutes. Unless otherwise stated samples to be affinity purified had 1% (v/v) Triton X-100 added. The samples were pooled and centrifuged at 7740g at 4°C for 10 minutes. The supernatant and pellet were snap frozen separately and stored at -70°C. When the pellet was shown to contain the recombinant protein the inclusion bodies were solubilised.

2.3.12.4 Inclusion Body Solubilisation

The pellet was resupended in 40 ml of buffer 1 [50 mM Tris (pH 8.0), 1 mM EDTA (pH 8.0), 100 mM NaCl] and 1 mg/ml lysozyme. This was incubated at room temperature for 20 minutes before being centrifuged at 5,000g for 10 minutes. The supernatant was removed and stored at -20°C. The pellets were transferred to ice and resuspended in 40 ml of ice cold buffer 2 [50 mM Tris (pH 8.0), 1 mM EDTA (pH 8.0), 100 mM NaCl, 0.1% (v/v) sodium deoxycholate] and incubated for 10 minutes. 8 mM $MgCl_2$ and 10 μg/ml DNase I were added and incubation continued at 4°C until the viscosity disappeared. The inclusion bodies were removed by centrifugation at 10,000g for 10 minutes. The supernatant was removed and stored at -20°C. The pellets were washed by resuspending in 12 ml of buffer 3 [50 mM Tris (pH 8.0), 1 mM EDTA (pH 8.0), 100 mM NaCl, 1% (v/v) Nonidet® P-40]. Centrifugation was

again carried out at 10,000g for 10 minutes and the pellets washed in 12 ml of buffer 1. After further centrifugation the pellets were resuspended in 8 ml 8 M urea followed by dialysis against 1 x PBS.

2.3.12.4.1 Dialysis

Dialysis tubing was washed with distilled H_2O followed by dialysis buffer. One end was clipped shut and the solution added to the other end, before that too, was clipped shut. Dialysis was performed at 4°C in 500 ml of dialysis buffer which was stirred for 1 hour and 30 minutes. The buffer was changed twice during this time and then the sample was removed from the dialysis tubing and snap frozen on dry ice.

2.3.12.5 Affinity Purification of Recombinant Protein

Glutathione Sepharose® 4B beads were prepared as a 50% (w/v) slurry in MTPBS buffer according to the manufacturers instructions. 100 µl of 50% (w/v) slurry were mixed with 1.25 ml of sonicated cell lysate or 500 µl of solubilised pellet fraction in 1x PBS by rotating at room temperature for 30 minutes. The samples were centrifuged at 500g for 10 minutes. The supernatant was removed carefully and 50 µl snap frozen for analysis by SDS-PAGE. The beads were washed 4 times in 6 ml of MTPBS to remove non-specifically bound protein. Glutathione Sepharose® 4B beads to be used in the whole genome PCR screen (section 2.3.18) were resuspended in 33 mM Tris (pH 8.0), 80 mM NaCl, 6.7 mM DTT, 0.67 mM PMSF and 20% (v/v) glycerol to give a 50% (v/v) slurry and snap frozen. The recombinant proteins were eluted from the Glutathione Sepharose® 4B beads by the addition of 375 µl of elution buffer [50 mM Tris (pH 8.0), 120 mM NaCl, 10 mM DTT, 1 mM PMSF, 20 mM reduced glutathione] per 100 µl of protein-bound beads. The samples were rotated at room temperature for 15-30 minutes and centrifuged at 500g for 10 minutes. The eluants were snap frozen on dry ice in 500µl aliquots and the elution step repeated.

2.3.12.6 Determination of Protein Concentration

The concentration of small aliquots of affinity purified recombinant protein was estimated by comparison to Coomassie Blue stained BSA standards which had been electrophoresed in a SDS-polyacrylamide gel.

2.3.12.7 Concentrating Proteins

Several methods were used in an attempt to concentrate the affinity purified recombinant proteins:

a) 2 ml of affinity purified protein was reduced to 200 µl using a Centricon 10 membrane unit, according to the manufacturer's instructions.

b) An equal volume of cold 100% (v/v) ethanol was added to affinity purified protein and incubated at -20°C for 30 minutes. The sample was centrifuged at 10,000g for 30 minutes at 4°C and the supernatant removed. The pellet was resupended in a suitable volume of buffer and snap frozen immediately.

c) An equal volume of cold 100% (v/v) acetone was added to affinity purified protein and incubated at -20°C for 30 minutes. The sample was centrifuged at 10,000g for 30 minutes at 4°C and the supernatant removed. The pellet was resupended in a suitable volume of buffer and snap frozen immediately.

2.3.12.8 Thrombin Cleavage of Recombinant Proteins

In an attempt to remove the GST-tag from the recombinant protein, various conditions were tried, to cleave the protein at a thrombin site present between the GST and the expressed protein of interest:

a) 0.01U thrombin were mixed with 1.0 µg of affinity purified protein in a total volume of 200 µl and incubated at room temperature for 24 hours. 15 µl aliquots were taken for SDS-PAGE analysis every 2 hours.

b) 0.05U thrombin were mixed with 5.0 µg of affinity purified protein in a total volume of 200µl and incubated at room temperature for 24 hours. 15 µl aliquots were taken for SDS-PAGE analysis every 2 hours.

c) 0.10U thrombin were mixed with 5.0 µg of affinity purified protein in a total volume of 200 µl and incubated at room temperature for 24 hours. 15 µl aliquots were taken for SDS-PAGE analysis every 2 hours.

2.3.13 SDS-Polyacrylamide Gel Electrophoresis

Proteins were analysed using dissociating, discontinuous SDS-PAGE (Laemmli, 1970). Electrophoresis was carried out in mini separating gels containing 10% or 15% (w/v) acrylamide/bis (29:1) in a BIO-RAD MiniProtean II gel apparatus. 0.75 mm thick mini gels with 15 wells were used. Separating gels were overlaid with 4% (w/v)

acrylamide/bis (29:1) stacking gels. Protein samples were mixed with 5 x or 2 x SDS-PAGE loading buffer, boiled for 10 minutes and electrophoresed in 1 x SDS-PAGE running buffer at 100V, 50mA until the dye band just left the bottom of the gel. 6 μg of low molecular weight marker proteins (Pharmacia Biotech) was electrophoresed in analytical gels and 16 μg of kaleidoscope prestained protein standard was also electrophoresed in gels that were electroblotted onto Hybond™-N ECL nitrocellulose membranes.

2.3.14 Protein Detection Methods

2.3.14.1 Coomassie Blue Staining

SDS polyacrylamide gels were stained with 0.1% (w/v) PhastGel Blue R (Coomassie R 350 stain) in 30% (v/v) methanol, 10% (v/v) acetic acid for 15 minutes, followed by destaining in 20% (v/v) methanol, 5% (v/v) acetic acid overnight.

2.3.14.2 Silver Staining

SDS polyacrylamide gels were prefixed in 50% (v/v) methanol, 10% (v/v) acetic acid for 30 minutes, followed by 5% (v/v) methanol, 7% (v/v) acetic acid for 30 minutes. The proteins were fixed in 10% (w/v) gluteraldehyde for 20 minutes and washed in distilled H_2O overnight. Gels were stained in 0.1% (w/v) $AgNO_3$ for 15 minutes, developed in 0.27 M Na_2CO_3, 0.05% (v/v) formaldehyde until bands were clearly visible. The reaction was stopped by adding 2 g of citric acid crystals. The stained gels were stored prior to drying in 10% (v/v) ethanol, 5% (v/v) glycerol.

2.3.15 Storage of SDS Polyacrylamide gels

Gels were stored by drying under vacuum at 80°C on a BIO-RAD Model 583 Gel dryer, onto 3 MM grade chromatography paper.

2.3.16 Western Blot Analysis

Recombinant proteins to be investigated by western analysis were separated by SDS-PAGE. Gels, Hybond™-N ECL nitrocellulose membrane and thick absorbent filter paper were equilibrated in western blot transfer buffer (25 mM Tris, 192 mM glycine, 20% (v/v) methanol) for 15 minutes. Proteins were transferred from the

polyacrylamide gel to the membrane with the BIO-RAD Trans-Blot® semi dry transfer cell for 60 minutes at 10-15V with maximum current.

Membranes were blocked in PBS-T/5% (w/v) skimmed milk powder at room temperature for 1 hour. The membranes were washed in PBS-T at room temperature for 30 minutes, with the PBS-T being changed every 5 minutes. The primary antibody was diluted (1/1000) in 5 ml of PBS/5% (w/v) skimmed milk powder. Diluted antibody was added to the membrane and incubated overnight at 4°C. The membranes were washed as previously in PBS-T and the secondary HRP-conjugated antibody was diluted (1/1000 for donkey anti-rabbit HRP-conjugated serum or 1/10000 for donkey anti-goat HRP-conjugated serum) in 5 ml of PBS/5% (w/v) skimmed milk powder. The diluted secondary antibody was added and incubated at room temperature for 1 hour before being washed as previously. To specifically detect proteins containing the GST-tag, a polyclonal goat anti-GST serum was used as the primary antibody with polyclonal donkey anti-goat IgGs conjugated with HRP as the secondary antibody. However, to confirm that proteins were DMAHP recombinant proteins, R254 DMAHP antiserum was used as the primary antibody with polyclonal donkey anti-rabbit IgGs conjugated with HRP as the secondary antibody. R254 DMAHP antiserum contains GST depleted polyclonal antibodies that were raised against a GST-DMAHP recombinant protein containing the homeodomain and 17 amino acids from the 3' end of the Six domain

Detection reagents from the ECL™ western analysis system (Amersham International plc.) were used according to the manufacturer's protocol to detect the signal produced by the antigen-antibody-antibody interactions. Detection reagent 1 and detection reagent 2 were mixed in an equal volume so that the final volume was 0.125 ml/cm^2 of membrane. The membranes were covered with the detection solution for 1 minute at room temperature and after removal of excess detection solution they were sealed in Saran® wrap and exposed to X-ray film in the presence of an intensifying screen. Autoradiographs were developed after 15 seconds to 20 minutes.

2.3.17 Gel Retardation Assays

2.3.17.1 Subcloning of DNA Probes

40 bp double stranded oligonucleotides representing putative DNA binding sites were created by annealing two single stranded oligonucleotides to give a sequence with a *Bam*H I half site at one end and a *Eco*R I half site at the other end. The oligonucleotides used are listed in table 2.4. ARE1 and ARE2 represent the ARE [from -96 to -62 / Suzuki-Yagawa, 1992 #4664] and include the 13 bp AREC3

binding site and 21 bp of flanking sequence. ARE-like1 and ARE-like2 (nucleotides 726-760, accession number L08835) represent the ARE-like sequence and GGATTA1 and GGATTA2 (nucleotides 1311-1346, accession number L08835) the GGATTA consensus site present in the promoter region of *DMPK* . A mutant ARE sequence that has been shown not to bind AREC3 was also produced (oligonucleotides, AREmut1 and AREmut2) (Kawakami *et al.*, 1996a). The double stranded oligonucleotides were subcloned into the *Bam*H I /*Eco*R I sites of pBluescript® SK(+) phagemid.

2.3.17.2 End-labelling of DNA Fragments

1-2 μg plasmid DNA, containing the putative DNA binding site, was digested with one or two restriction endonucleases, in a 20 μl volume, to produce a fragment containing at least one 5' overhang, consisting of at least one guanine. 20 μCi [α-^{32}P] dCTP (3000 Ci/nmol), 0.2 mM dNTPs(-dCTP) and 2U Klenow were added and incubated at room temperature for 20 minutes. The DNA was purified by phenol/chloroform extraction, ethanol precipitated and resuspended in 20 μl 1 mM Tris (pH 8.0).

2.3.17.3 Binding Reaction

7% (v/v) glycerol, 0.01 μg/μl poly[dI/dC].poly[dI/dC], 10 μl affinity purified recombinant protein or unpurified cell lysate diluted in 1.5 x binding buffer, 3.33x10^{-3}% (w/v) bromophenol blue, 100 cps radiolabelled probe and H_2O to bring the total volume to 15 μl were mixed on ice. They were incubated for 20 minutes at room temperature.

2.3.17.4 Non-denaturing Polyacrylamide Gel Electrophoresis

Binding reactions were analysed using non-denaturing PAGE. Electrophoresis was carried out in a BIO-RAD Protean II gel apparatus. 4% (w/v) or 8% (w/v) non-denaturing acrylamide/bis (37.5:1) gels were used (low ionic strength or 1 x Tris/glycine). The monomeric acrylamide gel mix was degassed for 15 minutes prior to the addition of the polymerising agents, APS and TEMED. Samples were electrophoresed in a pre-run gel for 3 hours at 200V, 32mA with continuous re-circulation of the appropriate non-denaturing polyacrylamide gel running buffer at 4°C. The gel was dried under vacuum at 80°C on a BIO-RAD Model 583 Gel dryer, onto 3 MM grade chromatography paper and exposed to X-ray film in the presence of an intensifying screen at -70°C.

2.3.18 Whole Genome PCR Screen

2.3.18.1 Preparation of DNA

1 μg of high molecular weight human lymphocyte DNA was cleaved to completion with either *Sau*3A I or *Tsp*509 I. The DNA cleaved with *Sau3*A I was then ligated to linkers consisting of 5μg of oligonucleotides V-SU and V-NL annealed together and the DNA cleaved with *Tsp*509 I was ligated to linkers consisting of 5 μg of oligonucleotides TSP and V-NL annealed together. The DNA was heated to 95°C, precipitated using 1.3 volumes of propan-2-ol and 0.3 volumes of 10 M ammonium acetate air dried and resuspended in 10 μl H_2O.

2.3.18.2 Optimisation of Binding Reaction

To determine the optimum length of time that the DNA needs to bind to the GST-Six+HD, Glutathione coated Sepharose® 4B beads a time course was performed. 1,000 cps of radiolabelled ARE fragment and pBluescript® SK(+) were mixed with either 200 or 500 μl of a 50% (v/v) slurry of GST-Six+HD coated Sepharose® 4B beads (about 4 or 10 μg of protein) or GST coated Sepharose® 4B beads (about 8 or 20 μg of protein) in an eppendorf tube. Three tubes were prepared for each protein. Samples were rotated at room temperature for either 15 minutes, 30 minutes or 1 hour before being centrifuged at 10,000g for 5 minutes and the supernatant removed. The specific activity of both beads and supernatant were measured. The beads were washed in 1 x DNA-protein binding buffer and the specific activity of beads and wash buffer were measured. DNA was eluted by the addition of 100 μl of 1x DNA-protein binding buffer plus 1M NaCl. The samples were rotated at room temperature for 30 minutes before being centrifuged at 10,000g for 5 minutes and the eluant removed. Again the specific activity of eluant and beads was measured.

2.3.18.3 Binding Reaction

300 ng of linkered genomic DNA (either *Sau*3A I cleaved or *Tsp*509 I cleaved) was mixed with 200 μl of a 50% (v/v) slurry of GST-Six+HD, Glutathione coated Sepharose® 4B beads (about 4μg of protein) in an eppendorf tube. The sample was rotated at room temperature for 1 hour before being centrifuged at 10,000g for 5 minutes and the supernatant removed. The beads were washed in 1 ml of 1 x DNA-protein binding buffer. Bound DNA was then eluted by the addition of 100 μl 1x DNA-protein binding buffer plus 1 M NaCl. The sample was rotated at room

temperature for 30 minutes before being centrifuged at 10,000g for 5 minutes and the eluant removed. The DNA was precipitated by the addition of 1.3 volumes of propan-2-ol, air dried and resuspended in 10µl H_2O.

2.3.18.4 Optimisation of PCR Conditions

PCR conditions were optimised using unbound linkered genomic DNA as a template. The basic conditions are described in section 2.3.6. The primer utilised was 0.5 µM of V-NL. Perkin Elmer AmpliTaq Gold™was used and as this enzyme is thermostable a 12 minute 94°C pre-PCR stage was added to the basic cycle to provide an efficient hot start. Annealing temperatures of 55°C and 59°C were tried in conjunction with 1 mM and 2 mM spermidine and 1% -7% (v/v) DMSO. Reactions were performed on Gene Amp® PCR System 9700

2.3.18.5 PCR

PCR was carried out using Perkin Elmer AmpliTaq Gold™, 0.5 µM of primer V-NL, an annealing temperature of 59°C and 5% DMSO. The templates used were 5 µl of eluted, bound and precipitated DNA , 5 µl of 1/10 and with later rounds 1/100 and 1/1000 dilutions of the DNA. 12 µl of PCR reaction was electrophoresed in a 1% (w/v) agarose gel. 100-400 ng of PCR product was used in another round of selection.

2.3.18.6 Subcloning of PCR products

Rounds of binding and PCR were performed until discrete bands could be seen on an agarose gel. At this stage PCR products were purified by adapting the Qiagen gel extraction kit protocol. 60 µl of PCR product, from multiple reactions was mixed with 180 µl of solution QG from the kit. The manufacturer's instructions were then followed with the omission of the heating step required to melt agarose. Purified DNA was then subcloned into pGEM®-T Easy vector following the manufacturer's instructions. Clones were sequenced using T7 and SP6 primers.

2.3.19 DNase I Footprinting

2.3.19.1 End-labelling of DNA Fragment

A DNA fragment for DNase I footprinting must be labelled at only one end. Therefore, 1-2 μg plasmid DNA was digested with 2 restriction endonucleases to produce a fragment containing only one 5' overhang, consisting of at least one guanine. For each fragment to be labelled, 4 identical restriction digests were performed. The restriction digests were electrophoresed in a 1% (w/v) agarose gel and the fragments cut out of the gel, pooled and purified using a Qiagen gel extraction kit. The DNA was eluted in 32 μl 1 mM Tris (pH 8.0). 10 μl DNA, 25 μCi [α-^{32}P] dCTP (3,000 Ci/nmol), 0.25 mM dNTPs(-dCTP), 4U Klenow, 1 x Life Technologies Ltd React 2 in a total volume of 20 μl were incubated at room temperature for 25 minutes. 0.25 mM dCTP was added and incubated for a further 5 minutes. The DNA was then purified by phenol/chloroform extraction, ethanol precipitated, air dried and resuspended in 20 μl 1 mM Tris (pH 8.0).

2.3.19.2 DNase I reaction

7% (v/v) glycerol, 0.01 μg/μl poly[dI/dC].poly[dI/dC], 10 μl of 1.5 x DNA-protein binding buffer or unpurified GST-Six+HD cell lysate or GST cell lysate, diluted 1/10 or 1/20 in 1.5 x DNA-protein binding buffer, 200 cps radiolabelled probe and H_2O to bring the total volume to 15 μl were mixed on ice. They were then incubated for 20 minutes at room temperature, before being divided into 3, 4 μl aliquots. 1 mg/ml DNase I stored in DNase I dilution buffer was diluted to 12.50 μg/ml, 6.25 μg/ml and 3.13 μg/ml in DNase I dilution buffer plus 50 mM $MgCl_2$. 1 μl of diluted DNase I was added to each of the three tubes, such that all samples containing each protein dilution were digested by each dilution of DNase I and incubated at room temperature for exactly 5 minutes. A no DNase I control was carried out on a fourth 1.5 x DNA-protein binding buffer tube. 5 μl of DNase I stop solution was then added to each tube. Restriction digests of the radiolabelled DNA probe were performed, to create size markers. 5 μl of each digest was removed and placed in separate tubes to which 5 μl of DNase I stop solution was added. All samples were heated to 90°C for 2 minutes before being cooled on ice.

2.3.19.3 Denaturing Polyacrylamide Gel Electrophoresis

DNase I footprinting reactions were analysed in a 6% (w/v) acrylamide/bis (19:1) manual sequencing gel. BIO-RAD AG 501-XB (D) resin (20-50 mesh) was added to

the urea, acrylamide and water to de-ionise the mix. This mix, together with the TBE was degassed for 10 minutes before the addition of the polymerising agents, APS and TEMED. 5 µl of each reaction was loaded onto a pre-run gel and electrophoresed for 2 hours 10 minutes at a constant power of 40W. The gel was dried under vacuum at 80°C on a BIO-RAD Model 583 Gel dryer, onto 3 MM grade chromatography paper and exposed to X-ray film in the presence of an intensifying screen at -70°C.

Chapter

3

Screening for a Human *DMAHP* cDNA Clone

3.1 Introduction

The aim of the work presented in this chapter was to isolate a human *DMAHP* cDNA, that could be cloned into an expression vector, for the production of recombinant protein. The recombinant protein could then be utilised in DNA-protein binding studies to investigate the DNA binding properties of DMAHP. Prior to the start of this project over 6 million independent cDNA clones had been screened by conventional hybridisation methods and no *DMAHP* cDNA identified. These were from mouse heart (King, 1996), human foetal brain, human adrenal gland, human foetal muscle, human adult muscle (Boucher, 1996), and human foetal muscle (Winchester, 1997) cDNA libraries. Therefore, it was decided to screen libraries using a small pool PCR-based method, which is more sensitive to rare DNA sequences (Israel, 1993).

3.2 Optimisation of PCR

PCR primers, DMAHPFB and DMAHPRC were specifically designed to reduce the likelihood of DNA from sources other than the cDNA libraries interfering with results. DMAHPFB is situated in exon B and DMAHPRC is in exon C, downstream of any cloned RT-PCR products present in the laboratory (figure 3.1). Therefore, genomic PCR products can be distinguished from cDNA PCR products and cloned RT-PCR products will not amplify. The primers are also situated outside of any region of homology to other known members of the *Six* subfamily.

200 ng of human genomic lymphocyte DNA was used as a template for the optimisation of PCR conditions. The basic PCR conditions are described in section 2.3.6. An annealing temperature of 57°C and a range of DMSO concentrations from 0-5% (v/v) were used. Each set of reaction conditions was tested using *Taq* polymerase supplied by Boehringer Mannheim and Promega (storage buffer A) using buffer supplied with the enzyme. 70% (v/v) of each PCR product was electrophoresed in a 2% (w/v) agarose gel (figure 3.2a). 4% (v/v) DMSO and Boehringer Mannheim *Taq* polymerase were chosen as the optimum conditions. As multiple bands were present, the annealing temperature was raised to 59°C and this resulted in the amplification of a single product of the expected size (708 bp) (figure 3.2b).

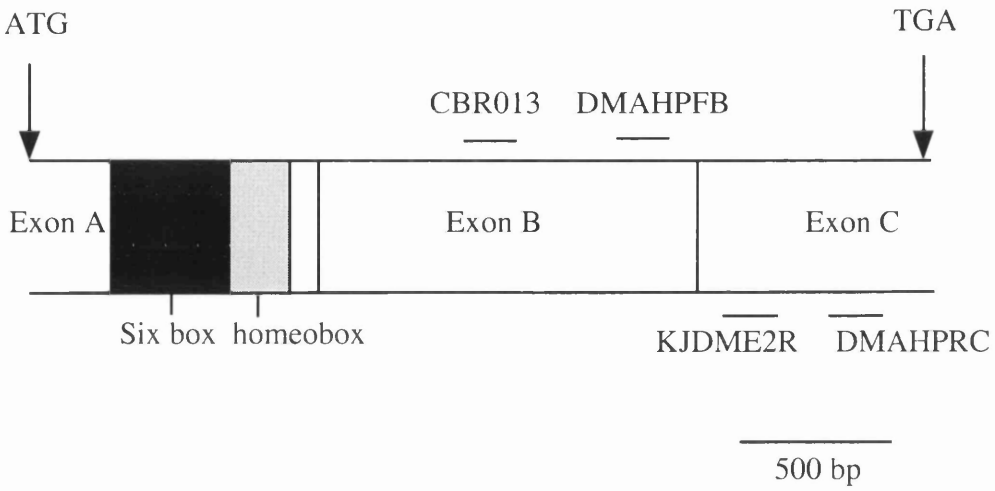

Figure 3.1 Structure of *DMAHP* showing PCR primers

A schematic diagram of *DMAHP* indicating the positions of the primers DMAHPFB, CBR013, KJDME2R and DMAHPRC, used to screen cDNA libraries for *DMAHP*. The predicted translation start site (ATG) and the stop site for the ABC isoform (TGA) are shown.

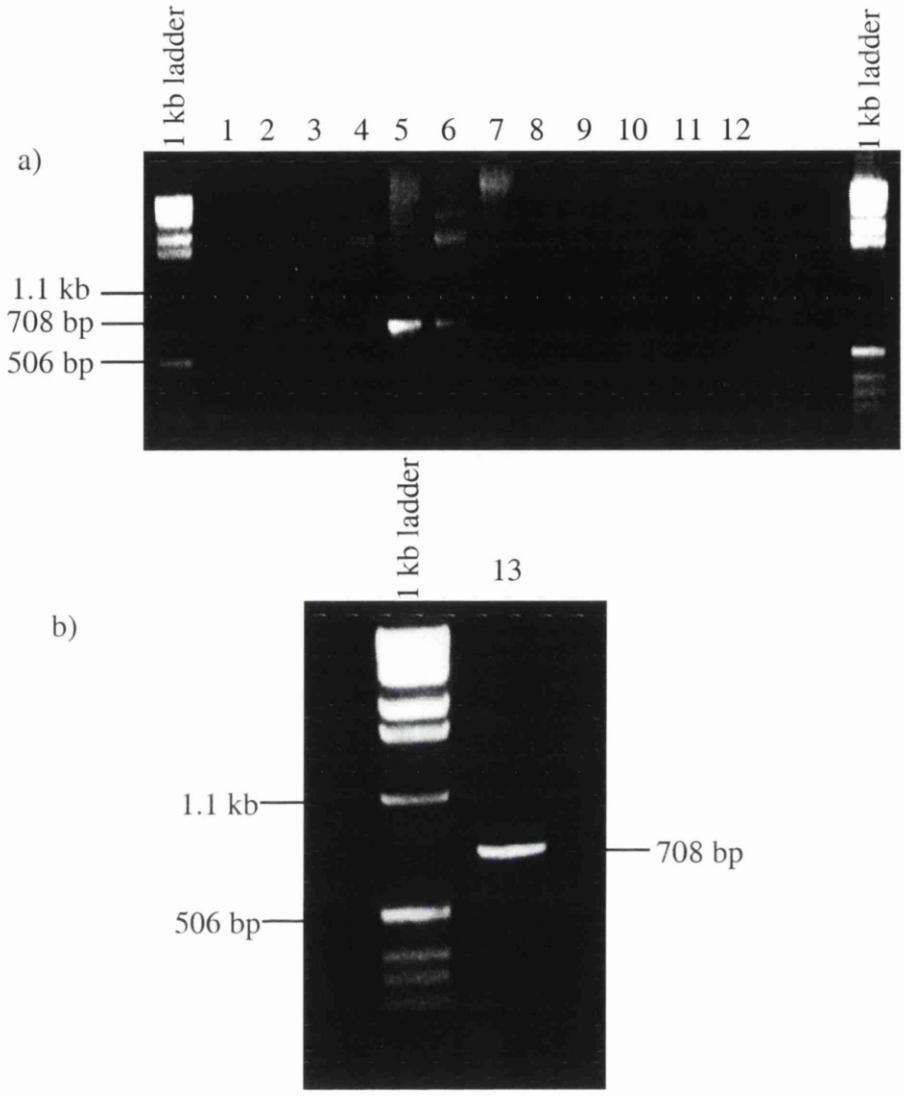

Figure 3.2 Analysis of PCR optimisation

Human genomic DNA was amplified using primers DMAHPFB and DMAHPRC. 70% (v/v) of each reaction was electrophoresed in a 2% (w/v) agarose gel. To optimise the PCR the following conditions were tried with an annealing temperature of 57°C: Boehringer Mannheim *Taq* polymerase and DMSO concentrations, 0, 1, 2, 3, 4 and 5% (v/v) DMSO (lanes 1-6 respectively), Promega *Taq* polymerase and DMSO concentrations, 0, 1, 2, 3, 4 and 5% (v/v) DMSO (lanes 7-12 respectively). Boehringer Mannheim *Taq* polymerase, 4% (v/v) DMSO and an annealing temperature of 59°C resulted in a single band of 708 bp (lane 13).

3.3 PCR-based screening of cDNA Libraries

PCR using optimised conditions (section 3.2) was performed on 10 ng, 1 ng and 0.1 ng of various plasmid cDNA libraries (table 3.1). Brain and muscle cDNA libraries were chosen as RT-PCR analysis demonstrated that *Six5/Dmahp* is expressed in these tissues in mouse (King, 1996).

tissue	vector	method of production
human foetal brain	pcDNAII	oligo dt primed
human foetal brain	pcDM8	oligo dt primed
human foetal muscle	pcDM8	oligo dt primed
human foetal brain	pcDM8	random hexamer primed

Table 3.1 cDNA libraries screened by PCR

70% (v/v) of each PCR was electrophoresed in a 2% (w/v) agarose gel. As no DNA was visible on the agarose gel when viewed under ultra violet light, it was transferred to Hybond™-N membrane. 1.8S2 (a gift from Catherine Boucher), a pBluescript® SK(+) clone containing a 1.8 kb genomic *DMAHP Sac* II fragment covering the region amplified by the PCR, was digested with *Sac* II. The excised fragment was cut out of a 1% (w/v) LMP agarose gel and purified with hot phenol, before being radiolabelled. The radiolabelled probe was used to screen the Hybond™-N membrane for PCR products. The human foetal brain cDNA library in vector pcDNAII produced a band of the correct size (470 bp), which reduced in intensity as the amount of template DNA decreased. A faint band, of 470 bp, was also visible in the 10 ng random hexamer primed human foetal brain cDNA library in vector pcDM8 (figure 3.3). These results indicate the presence of a *DMAHP* cDNA in these two libraries. As the human foetal brain cDNA library in vector pcDNAII library gave the strongest bands, it was decided to continue with this library.

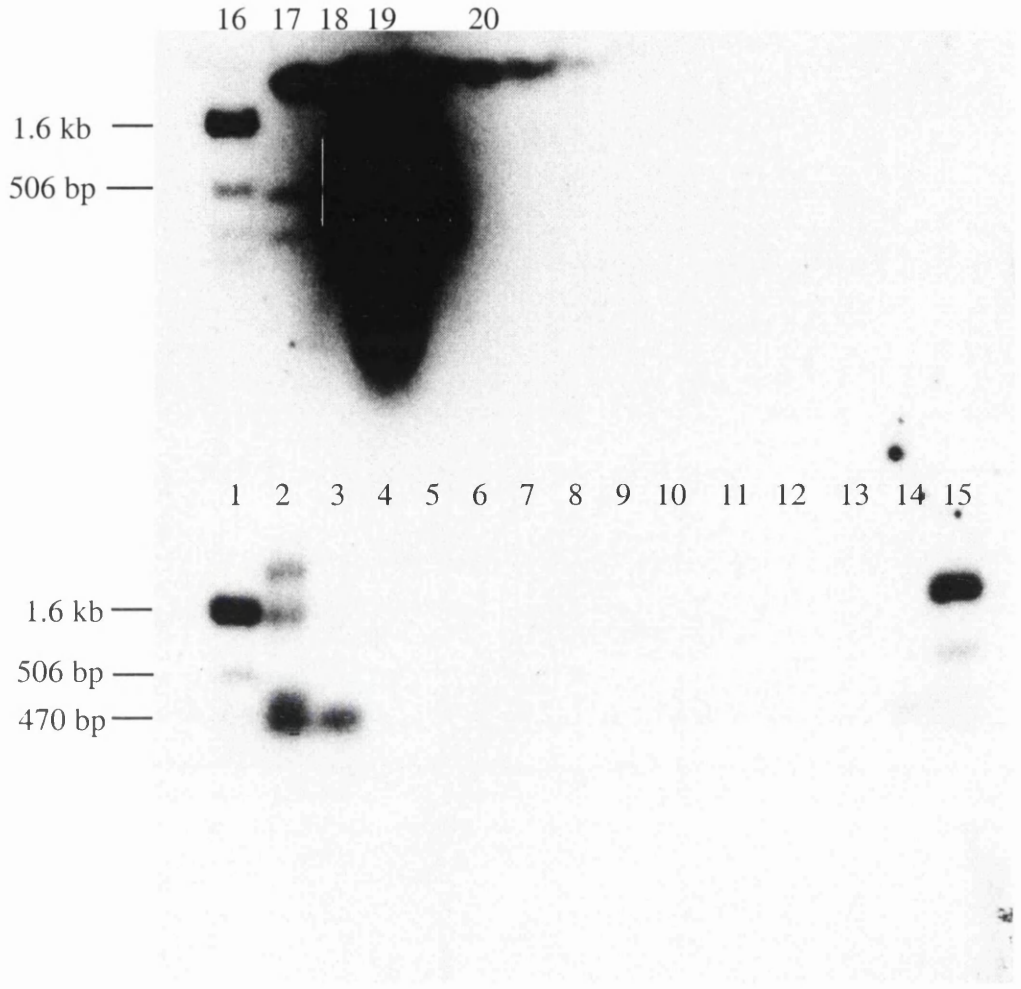

Figure 3.3 Southern blot analysis of an initial PCR-based library screen

A range of concentrations of 4 plasmid cDNA libraries was amplified using primers DMAHPFB and DMAHPRC. 70% (v/v) of each reaction was electrophoresed in a 2% (w/v) agarose gel. The gel was Southern blotted and probed with 1.8S2.
The lanes are as follows: 1 kb ladder (lane 1), human foetal brain (vector pcDNAII), 10 ng (lane 2), 1 ng (lane 3), 0.1 ng (lane 4), oligo dt primed foetal brain (vector pcDM8), 10 ng (lane 5), 1ng (lane 6), 0.1 ng (lane 7), oligo dt primed foetal muscle (vector pcDM8), 10 ng (lane 8), 1 ng (lane 9), 0.1 ng (lane 10), pcDM8, 10 ng (lane 11), 1 ng (lane 12), 0.1 ng (lane 13), random hexamer primed foetal brain (vector pcDM8), 10 ng (lane 14), 1 kb ladder (lanes 15 and 16), random hexamer primed foetal brain (vector pcDM8), 1 ng (lane 17), 0.1 ng (lane 18), genomic DNA, 200 ng (lane 19) and H_2O (lane 20).

3.4 Cloning and Sequencing of PCR product

To confirm that the PCR product was indeed *DMAHP*, the PCR was repeated on 10 ng of the library. 70% (v/v) of the PCR was electrophoresed in a 2% (w/v) agarose gel and a slice was excised from the gel at the position where the 470 bp PCR product was expected. 15 µl of H_2O was added to the slice and vortexed before being frozen (-70°C) and thawed (by incubation at room temperature) three times. 1µl was used as a template for 6 identical PCRs to re-amplify the DNA present in the gel slice. Two major bands of between 4 and 500 bp were amplified (figure 3.4). The bands were excised from the gel, identical sized bands pooled and the DNA purified using a Qiagen gel extraction kit. The DNA was subcloned into the T-vector, pMOS*Blue* (following manufacturer's instructions) and sequenced with the T7 primer. The sequences were compared to the GenBank, EMBL, DDBJ and PDB databases using the BLASTN program at NCBI. This confirmed that the larger product was amplified from *DMAHP* and was 470 bp in length and the smaller product was sequence from an unknown source and was presumably amplified as the result of the primers annealing non-specifically. The 470 bp *DMAHP* PCR product clone will now be referred to as DMAHP-470.

3.5 Small Pool PCR-Based Screening of a cDNA Library

XL-1 BLUE competent cells were transformed with 1 x 10^6 clones from the human foetal brain cDNA library in pcDNAII vector. This library was originally amplified from 3 x 10^6 clones (Derek Blake, personal communication). The cells were grown in an 8 x 8 matrix in a microtitre plate and then pooled to limit the number of samples for analysis to 16 (section 2.3.11.1). Plasmid DNA was extracted from each pool, diluted 1/100 and amplified by primers DMAHPFB and DMAHPRC using optimised PCR conditions (section 3.2). 70% (v/v) of each PCR was electrophoresed in a 2% (w/v) agarose gel. No bands were visualised on the ethidium bromide stained agarose gel with the exception of the positive control, which used 200 ng of genomic DNA as a template, giving a 708 bp product. Therefore, the DNA was transferred to Hybond™-N membrane and hybridised to radiolabelled DMAHP-470. Seven pools contained the 470 bp DMAHP PCR product, indicating 12 possible positive wells (figure 3.5a). Several pools either contained genomic clones or were contaminated with genomic DNA as indicated by the presence of a 708 bp band. DNA from the 12 possible positive wells was prepared and 1 µl of a 1/100 dilution, used as a template for further PCR reactions. Again no bands were visualised on an ethidium bromide stained 2%

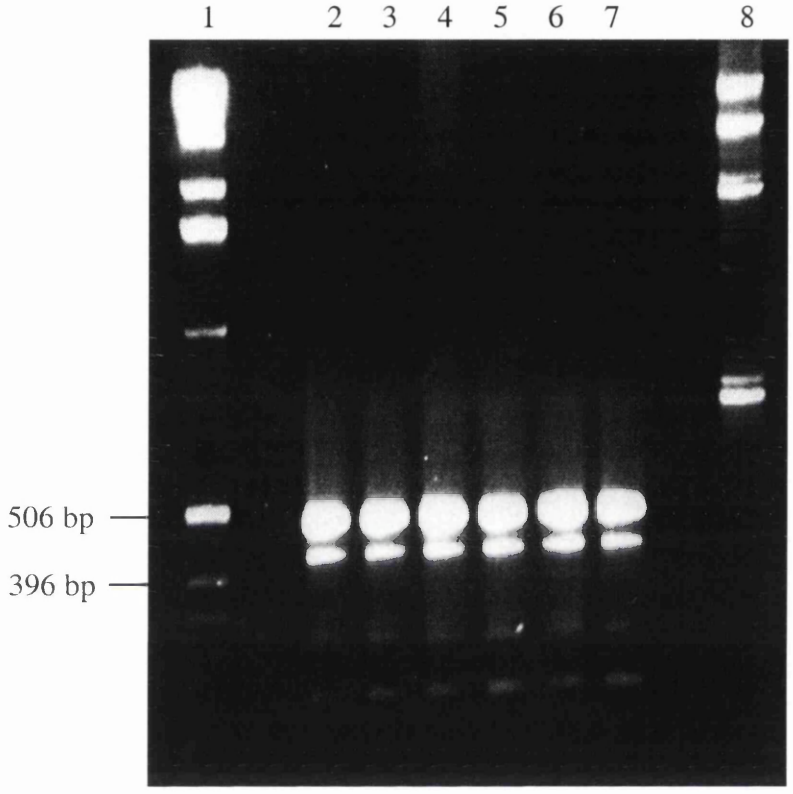

Figure 3.4 PCR amplification of human foetal brain cDNA library

Human foetal brain cDNA library was re-amplified using primers DMAHPFB and DMAHPRC. 70% (v/v) of each reaction was electrophoresed in a 2% (w/v) agarose gel. The lanes are as follows: 1 kb ladder (lane 1), re-amplified human foetal brain cDNA library (vector pcDNAII) (lanes 2-7), 200 ng human genomic DNA (lane 8).

Figure 3.5 Southern blot analysis of PCR-based library screen

PCRs were performed on 1 μl of a 1/100 dilution of plasmid DNA prepared from cultures of all the pools and appropriate wells at each stage of the library screening. 70% (v/v) of each PCR product was electrophoresed in a 2% (w/v) agarose gel. Gels were Southern blotted and probed with DMAHP-470.

a) First Round Pools

1 kb ladder (lanes 1, 15 and 16), pools 1-8 (lanes 2-9), pools A-H (lanes 10-14 and 17-19), 200 ng genomic DNA (lane 20), H_2O control (lane 21).

b) First Round Wells

1 kb ladder (lane 1), well 2A (lane 2), 2G (lane 3), 2H (lane 4), 6A (lane 5), 6E (lane 6), 6G (lane 7), 6H (lane 8), 7A (lane 9), 7E (lane 10), 7G (lane 11), 7H (lane 12), 200 ng genomic DNA (lane 13), H_2O control (lane 14).

c) Second Round Pools - made from 2A

1 kb ladder (lanes 1, 16-17), pools 1-8 (lanes 2-9), pools A-H (lanes 10-15 and 18-19), 200 ng genomic DNA (lane 20), H_2O control (lane 21).

d) Second Round Pools - made from 6H

1 kb ladder (lanes 1, 16-17), pools 1-8 (lanes 2-9), pools A-H (lanes 10-15 and 18-19), 200 ng genomic DNA (lane 20), H_2O control (lane 21).

e) Second Round Wells

from plate 2A

1 kb ladder (lane 1), well 1A (lane 2), 1B (lane 3), 1C (lane 4), 1D (lane 5), 1E (lane 6), 1H (lane 7), 2B (lane 8), 3B (lane 9), 5B (lane 10), 6B (lane 11), 7B (lane 12), 8B (lane 13), 200 ng genomic DNA (lane 13).

from plate 6H

1 kb ladder (lane 15), well 5B (lane 16), 5C (lane 17), 5F (lane 18), 5G (lane 19), 1C (lane 20), 2C (lane 21), 3C (lane 22), 6C (lane 23), 200 ng genomic DNA (lane 24), H_2O control (lane 25).

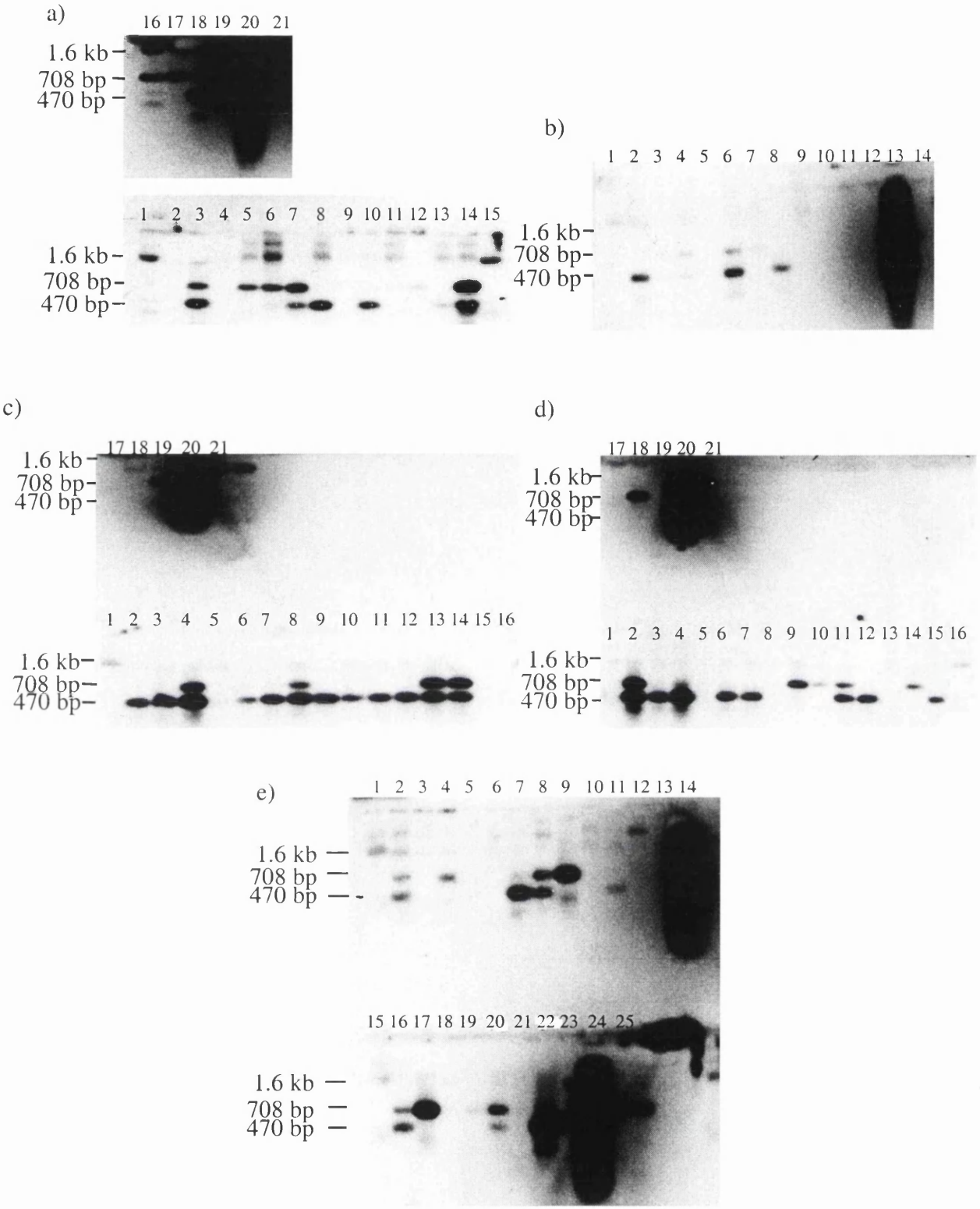

Figure 3.5 Southern blot analysis of PCR-based library screen

(w/v) agarose gel. However, when the gel was blotted and probed, three wells (2A, 6E and 6H) gave bands of 470 bp (figure 3.5b). 6E also contained genomic DNA and therefore it was decided to proceed with 2A and 6H. The number of clones in wells 2A and 6H was determined (section 2.3.11.1) to be 6 x 10^5 and l 2 x 10^3 clones/μl respectively.

40,000 clones from each of these wells were grown in an 8 x 8 matrix in a microtitre plate and then pooled. Plasmid DNA was prepared from the 16 pools, diluted 1/100 and amplified using primers DMAHPFB and DMAHPRC. 2% (w/v) agarose gels were once again blotted and the DNA hybridised to DMAHP-470. Plate 2A (figure 3.5c) contained 13 positive pools and plate 6H (figure 3.5d) contained 8 positive pools, indicated by 470 bp PCR products. Again some of the pools contained genomic DNA, indicated by 708 bp products. DNA from 12 of the possible positive wells from plate 2A and 8 of the possible positive wells from plate 6H was prepared and tested by PCR, blotting and hybridisation. Three wells from each plate (2A-1A, 2A-1H, 2A-2B, 6H-5B, 6H-1C and 6H-3C) contained the 470 bp *DMAHP* PCR product (figure 3.5e).

3.6 Determination of cDNA Insert Size

To determine the size of the inserts present in the positive clones within the three positive wells (2A, 6E and 6H), plasmid DNA was purified and 2 μg digested with *Eco*R I to release the inserts. The DNA was electrophoresed in a 1% (w/v) agarose gel and the DNA transferred to Hybond™-N membrane. The membrane was screened with radiolabelled DMAHP-470. Surprisingly, no positive inserts were identified (results not shown).

3.7 Screening of Clones from Positive Well

As 21 of the 32 pools from plates 2A and 6H in the previous round were positive for *DMAHP* it was decided to screen individual clones from a single positive well from one of these plates. The number of clones in well 6H-5B was determined (section 2.3.11.1) to be 4.7 x 10^3 clones/μl. 3,000 clones from 6H-5B were screened by hybridisation to radiolabelled oligonucleotide KJDME2R, which lies in the region amplified by DMAHPFB and DMAHPRC (figure 3.1), to replica filters (sections 2.3.11.2-2.3.11.4). No positive clones were identified (results not shown).

3.8 Confirmation of Positive PCR Results

As no positive clones were identified by hybridisation of *DMAHP* to the inserts released from the clones in wells 2A, 6E and 6H or by the screening of replica filters, it was hypothesised that some of the DNA preparations may have been contaminated with DMAHP-470. Therefore 10 ng of the library and the three positive wells (2A, 6E and 6H) from the first round of screening were re-screened, by PCR, using primers CBR013 (which lies 356 bp upstream of DMAHPFB, figure 3.1) and DMAHPRC. PCR conditions were optimised as section 3.2. No DMSO, an annealing temperature of 57°C and Boehringer Mannheim *Taq* polymerase were chosen as the optimal conditions. 70% (v/v) of each PCR product was electrophoresed in a 1.5% (w/v) agarose gel which was blotted and probed with DMAHP-470 (figure 3.6). Only 10 ng of library gave a product of the expected size (826 bp). These results indicate that some of the DNA preparations were indeed contaminated with cloned PCR product, DMAHP-470, and that the *DMAHP* cDNA identified in the human foetal brain cDNA library, was not present in the DNA prepared from the transformed XL-1 BLUE competent cells in wells 2A, 6E and 6H.

3.9 Direct Cloning of cDNA by Vector to Internal Primer PCR

As no *DMAHP* cDNA was present in the transformed XL-1 BLUE competent cells, direct cloning of the cDNA from the library, using vector and internal primers was attempted. The orientation of the insert within the vector was unknown. Figure 3.7 shows the two possible orientations and the primers that were chosen for this experiment. The idea was to perform PCR using vector and internal primers to produce two overlapping PCR products that could be cloned, separately into vectors and then ligated via a restriction site in the overlapping region, to produce the cDNA that is present in the library.

To determine the orientation of the insert, 8 PCRs were performed on 10 ng of human foetal brain cDNA library in pcDNAII vector using the following combinations of primers, T7-DMAHPFB, T7-DMAHPRC, T7-CBR013, T7-CBR005, SP6-DMAHPFB, SP6-DMAHPRC, SP6-CBR013, SP6-CBR005. DMAHPFB and DMAHPRC were used as a positive control. No DMSO, an annealing temperature of 55°C and Boehringer Mannheim *Taq* polymerase were used in the PCR. 70% (v/v) of each PCR product was electrophoresed in a 1% (w/v) agarose gel and the DNA transferred to Hybond™-N membrane. Radiolabelled DMAHP-470 was hybridised to the membrane. DMAHPFB-DMAHPRC amplified a cDNA product of 470 bp and also a genomic DNA product of 708 bp indicating the presence of genomic DNA in

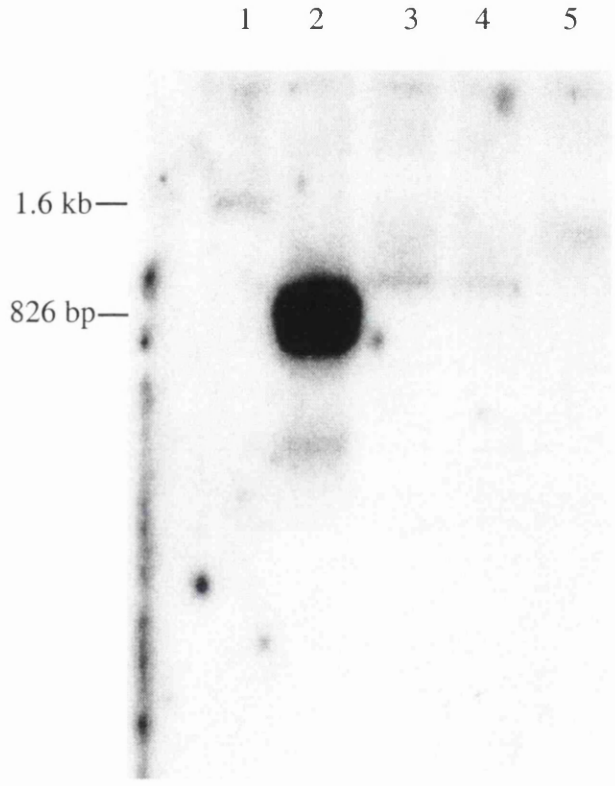

Figure 3.6 Southern blot analysis of PCR, using primers CBR013 and DMAHPRC

PCR was performed using primers CBR013 and DMAHPRC. 70% (v/v) of each reaction was electrophoresed in a 1.5% (w/v) agarose gel which was Southern blotted and probed with DMAHP-470. The lanes are as follows: 1 kb marker (lane 1), 10 ng human foetal brain cDNA library (lane 2), 1µl 1/10 dilution 2A (lane 3), 6E (lane 4) and 6H (lane 5).

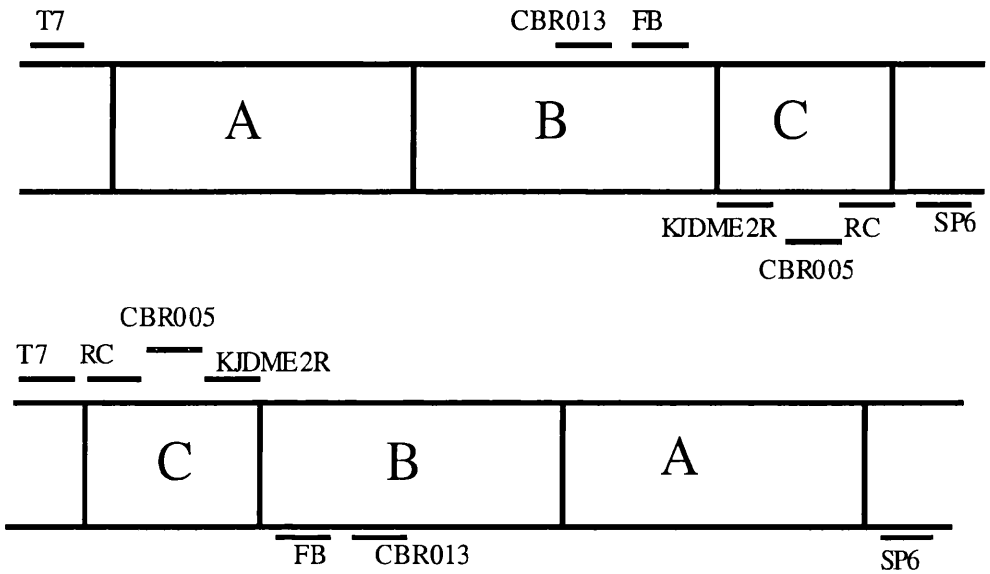

Figure 3.7 *DMAHP* and Primers

A schematic diagram of *DMAHP* showing the two possible orientations in pcDNAII and the positions of the internal *DMAHP* (DMAHPFB, DMAHPRC, CBR013, CBR005, KJDME2R) and vector (T7, SP6) primers.

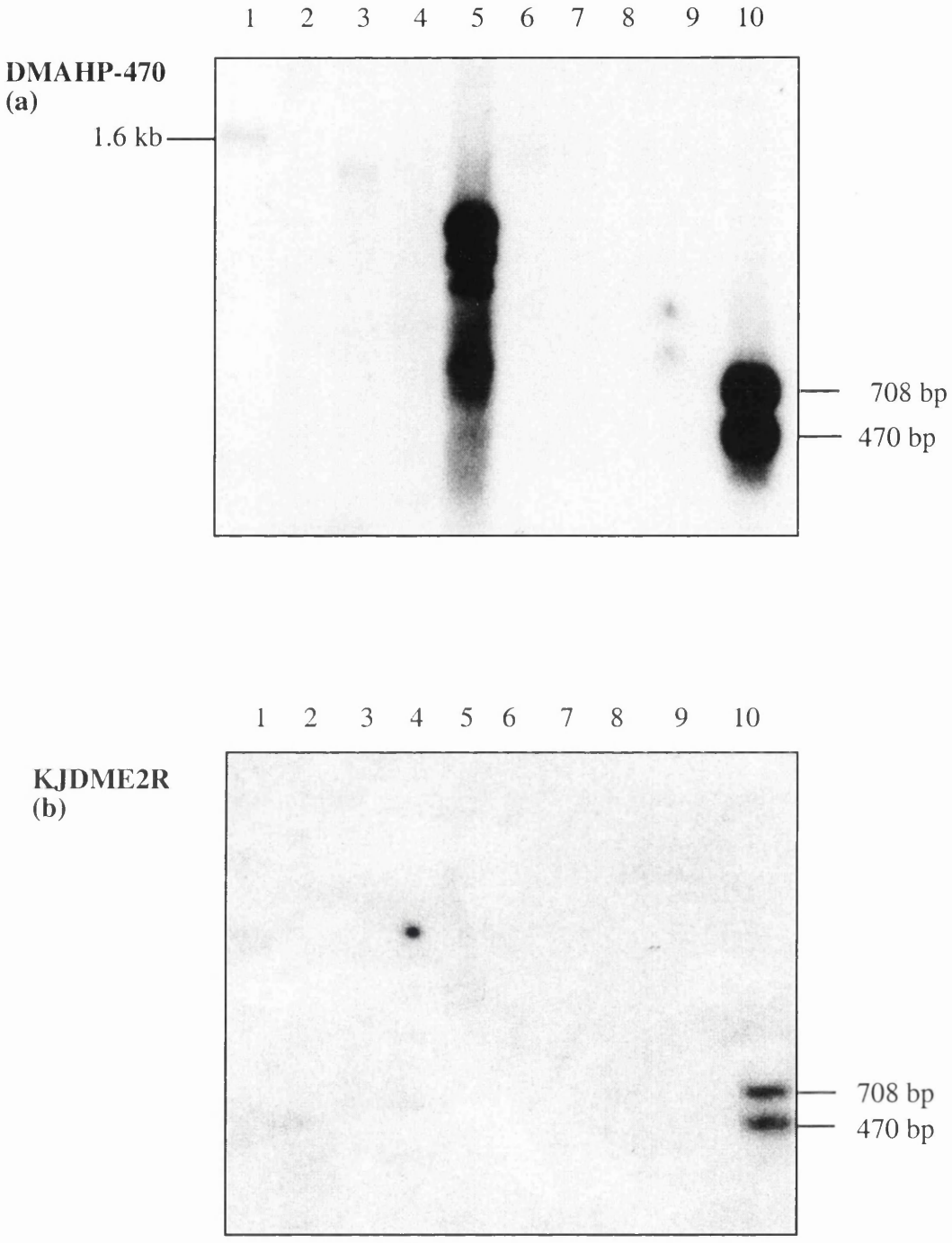

Figure 3.8 Southern blot analysis of direct cloning PCR

PCR was performed on 10 ng human foetal brain cDNA library using 8 different combinations of internal and vector primers (figure 3.8). 70% (v/v) of each reaction was electrophoresed in a 1% (w/v) agarose gel. The gel was Southern blotted and probed with DMAHP-470 (a), stripped and then probed with KJDME2R (b).The lanes are as follows: 1 kb ladder (lane 1), T7-DMAHPFB (lane 2), T7-DMAHPRC (lane 3), T7-CBR013 (lane 4), T7-CBR005 (lane 5), SP6-DMAHPFB (lane 6), SP6-DMAHPRC (lane 7), SP6-CBR013 (lane 8), SP6-CBR005 (lane 9), DMAHPFB-DMAHPRC (lane 10).

the reaction. Four bands were present in the T7-CBR005 lane (figure 3.8). To confirm that they were *DMAHP* the membrane was stripped and hybridised to a radiolabelled oligonucleotide KJDME2R, which lies in the region amplified by T7-CBR005. Disappointingly no bands were detected, with the exception of the two bands in the positive control lane (figure 3.8). The T7-CBR005 PCR was repeated and the four bands excised from an agarose gel. The DNA was purified using a Qiagen gel extraction kit and subcloned into the T-vector, pMOS*Blue* and sequenced. None of the clones contained *DMAHP*. However, two of the sequences were present in the GenBank database and when they were analysed they were shown to contain regions 100% homologous to the 10 most 3' bases of CBR005 although the rest of the sequence showed no homology to *DMAHP*. Regrettably, at this point, the search for a *DMAHP* cDNA was abandoned and an alternative method for the production of *DMAHP* expression constructs was devised (chapter 4).

3.10 Database searching

The GenBank, EMBL, DDBJ and PDB databases were regularly searched using the human genomic *DMAHP* sequence (accession number x848413) with the BLASTN program at NCBI to see if a cDNA was available from another source. One such database search revealed the presence of an EST which may be a *DMAHP* cDNA (accession number N23083). The clone was obtained and sequenced, but unfortunately contained a human haemoglobin gene. At some stage the clones were mislabelled and the distributors are now unable to supply the correct clone.

3.11 Discussion

The aim of the work described in this chapter was to isolate a full length human *DMAHP* cDNA. As conventional hybridisation techniques had previously failed to isolate such a cDNA, the more sensitive method of small pool PCR was undertaken. A 470 bp fragment of *DMAHP* was successfully amplified by PCR from a human foetal brain cDNA library and cloned. This fragment contains 322 bp of previously uncloned cDNA. Unfortunately the full sequence of this cDNA failed to be isolated.

XL-1 BLUE competent cells were transformed with 1 x 10^6 clones from the cDNA library. However, it appears that *DMAHP* was not present in the DNA prepared from cells, post-transformation. This may have been due to low levels of *DMAHP* clones in the cDNA library chosen and therefore competent cells with a higher transformation efficiency may be required when using this particular cDNA library, or the clone may

have been unsuitable for transformation. It is possible, that the *DMAHP* clone was present in the transformed cells, but leaky expression of *DMAHP* caused reduced growth of cells containing the positive clone and thus they were lost during the screening. The growth kinetics of cultures expressing *DMAHP*, in bacterial expression systems have been investigated and DMAHP expression has been shown to significantly reduce the growth rate of cultures in which it is being expressed (section 4.3 and Winchester, 1997).

The most likely explanation for the false positive results during the small pool PCR screen is that some of the preparations of DNA from the library pools and wells became contaminated with the cloned 470 bp PCR product. Alternatively the contamination occurred during the setting up of the PCRs. However, as the H_2O controls were always negative, this seems unlikely. Unfortunately, due to the sensitivity of the PCR, combined with the sensitivity of the hybridisation, this type of experiment will always be vulnerable to contamination problems. Future experiments should be performed under meticulous conditions designed to help prevent contamination, for example the DNA preparation and the PCR should be carried out away from any source of cloned DNA containing the primer sites. Contamination by human genomic DNA may also have occurred at one of these stages. Alternatively, the cDNA library contained genomic clones as contamination of cDNA libraries by genomic clones is known to be a problem when screening libraries. A previous cDNA library screen of a human foetal brain cDNA library (Stratagene) for a *DMAHP* cDNA clone, by conventional hybridisation led to the isolation of a genomic *DMAHP* clone (Boucher, 1996).

Direct cloning of the cDNA by PCR using vector to internal primers failed to amplify any *DMAHP* sequence. Of the sequences that were amplified, 2 are present in the GenBank database and were shown to contain regions 100% homologous to the 10 most 3' bases of the primer CBR005. This explains their amplification by PCR. The reason for their hybridising to DMAHP-470 is unclear, but is likely to be due non-specific hybridisation, to the relatively high concentrations of PCR amplified DNA present on the membrane. DMAHP-470 has a high G+C content and will therefore hybridise preferentially to other DNA sequences with a high G+C content. The 4 sequences that hybridised to DMAHP-470 all contained over 50% G+C. As oligonucleotide hybridisation tends to be more specific, KJDME2R would not have hybridised.

These results suggest that although a partial cDNA is present in the human foetal brain cDNA library (vector pcDNAII), it is either present at very low levels or is a rearranged clone. When this project was instigated, no human or murine *DMAHP*

cDNA had been isolated despite the screening of over 6 million independent clones. However, one extensively rearranged clone with additional mitochondrial DNA had been isolated from a mouse 5' stretch heart library (Catherine Boucher, personal communication). It is therefore likely, that *DMAHP* is not suitable for cloning in conventional systems and may be toxic to *E.coli*. Problems have been encountered when expressing DMAHP in bacterial expression systems including reduced growth rates of cultures expressing DMAHP (discussed above) failure to express any protein at all, the expression of truncated proteins and low levels of protein (section 4.3, Charles Thornton, personal communication and Winchester, 1997).

5' and 3' RACE have been attempted, to identify the 5' and 3' ends of *DMAHP* but have been unsuccessful (Catherine Boucher and Simon Carne personal communications). A partial mouse *Six5/Dmahp* has now been isolated from a mouse retina cDNA library (Kawakami *et al.*, 1996b) and therefore a human retina cDNA library has been screened using 5' and 3' *DMAHP* probes (Catherine Winchester, personal communication). Two truncated cDNA clones have been isolated, one a 5' clone and one a 3' clone. Five major transcription start sites of *DMAHP* have been identified by RNase protection analysis (Klesert *et al.*, 1997). Therefore, despite not being able to obtain a full length human *DMAHP* cDNA, the full structure of the gene is gradually emerging.

Chapter

4

Expression of Recombinant DMAHP Proteins

4.1 Introduction

The objectives of the work described in chapter 4 were to overexpress and affinity purify regions of DMAHP that could be used to characterise the DNA binding properties of this transcription factor. *DMAHP* shares a 528 bp region of sequence homology with other members of the *Six* subfamily of genes (section 1.8). This region encodes a homeodomain and immediately upstream of this, the Six domain (Oliver *et al.*, 1995a). At the time of starting this project nothing was known about the function of the Six domain. AREC3/Six4 was originally identified as one of the proteins that binds to the Na^+ K^+ ATPase α1 subunit regulatory element (ARE) (Suzuki-Yagawa *et al.*, 1992). It was shown that the homeodomain of murine AREC3 was required for DNA binding and the Six domain was required for specific binding to the ARE (Kawakami *et al.*, 1996a). To investigate the functions of these two domains in human DMAHP and to identify specific DNA binding sites, recombinant proteins encoded by the Six box and the homeobox were generated.

As no full length *DMAHP* cDNA was available, it was decided to subclone regions of the gene using pMOS-DMAHP-ABC (a 1.15 kb RT-PCR product, subclone, primed from DMAHPF and KJDME2R, figure 4.1) and cosmid F18894 (a 40 kb genomic clone containing *DMAHP* in its entirety). A Glutathione S-transferase (GST) Gene Fusion System (Pharmacia Biotech) was chosen for the expression and purification of the gene fragments as this system had previously been used with success in our laboratory (Winchester, 1997). It allows inducible overexpression of recombinant proteins as fusion proteins attached to the carboxyl terminus of GST. The recombinant proteins can be purified from bacterial cell lysates, if required, by affinity chromatography using Glutathione Sepharose® 4B.

4.2 Subcloning of the Six box and Homeobox into pGEX4T3

The Six box and the homeobox of human *DMAHP* were subcloned both separately and together, into pGEX4T3 for expression and purification. Figure 4.1 shows a schematic diagram of *DMAHP* and indicates the regions that were subcloned and expressed.

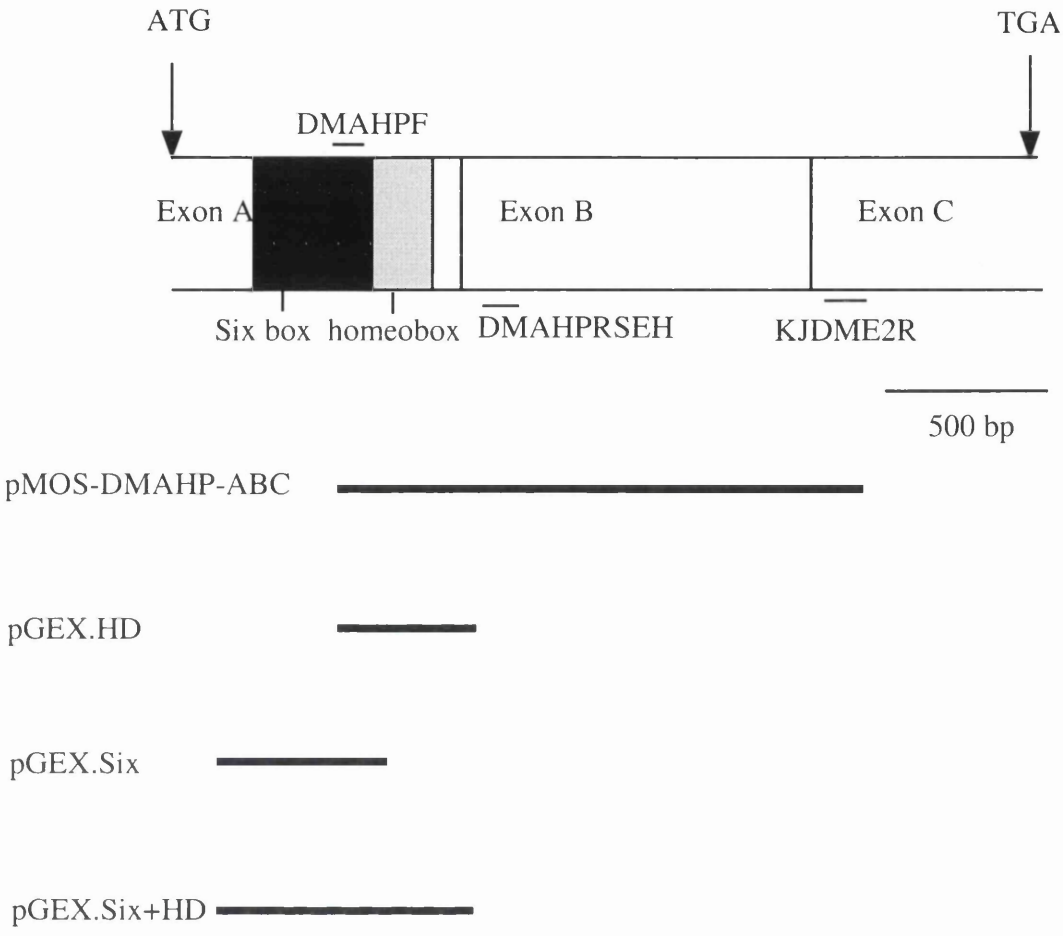

Figure 4.1 Structure of *DMAHP* showing subcloned regions

A schematic diagram of *DMAHP* indicating the region present in the RT-PCR product subclone, pMOS-DMAHP-ABC and the regions subcloned into pGEX4T3 for bacterial over expression (pGEX.HD, pGEX.Six and pGEX.Six+HD). PCR primers, DMAHPF, DMAHPRSEH and KJDME2R are indicated. The predicted translation start site (ATG) and the stop site for the ABC isoform (TGA) are shown.

4.2.1 Subcloning of the Homeobox into pGEX4T3

The homeobox (180 bp), 52 bp of the Six box and 79 bp of sequence 3' to the homeobox were amplified from pMOS-DMAHP-ABC by PCR using primers DMAHPF and DMAHPRSEH (figure 4.1) (section 2.3.6 with 1% (v/v) DMSO, Boehringer Mannheim Taq polymerase and an annealing temperature of 57°C). The 311 bp PCR product (figure 4.2a) was subcloned into the T-vector, pMOS*Blue* (pMOS.HD), following manufacturer's instructions. To confirm that the sequence had not been mutated by the PCR and to determine the orientation of the insert within pMOS*Blue*, pMOS.HD was sequenced using the T7 primer. No mutations were detected (results not shown) and the orientation of the insert is shown in figure 4.3. The orientation of the insert needed to be known to enable further subcloning.
As no suitable restriction endonuclease sites were present in pMOS*Blue* to allow the homeobox to be subcloned in frame into the expression vector pGEX4T3, the homeobox, from pMOS.HD, was subcloned into pBluescript® SK (+) (pSK (+).HD). This vector has an extensive MCS which is helpful when doing further subcloning (figure 4.4). pBluescript® SK(+) was digested with *Sma* I and CIAP treated to prevent it religating and the DNA extracted using phenol/chloroform. The PCR product containing the homeobox was released from pMOS.HD by digestion with *Bam*H I and *Hind* III and the polymerase activity of the Klenow fragment used to fill the recessed termini. The fragment was purified from a 3% (w/v) LMP agarose gel (figure 4.2b) using hot phenol and ligated into the *Sma* I site of pBluescript® SK(+). DH5α cells were transformed with the construct and 6 clones digested with *Acc* I to determine the orientation of the insert. Five of the 6 clones contained the insert in the same orientation. The sixth clone appeared to contain multiple inserts (figure 4.2d). Although the digest indicated that the 5 clones were in the same orientation both orientations would give a band of 272 bp and then a second band of either 114 bp or 55 bp depending on the orientation of the insert, which would electrophorese below the smallest visible band (134 bp) in the marker lane (figure 4.2d). However, as the 5 clones were identical it was impossible to determine which orientation they were in. Therefore, pSK(+).HD was sequenced with T7 and T3 (results not shown). This confirmed the identity and integrity of the insert and that the 5' end of the homeobox was at the *Eco*R I side of the MCS and the 3' end at the *Not* I side (figure 4.4).

Figure 4.2 Subcloning of the homeobox into pGEX4T3

Vector and insert DNA used in the subcloning of the homeobox into pGEX4T3 were analysed by electrophoresis in 1, 2 or 3% (w/v) agarose gels. LMP agarose gels were utilised when the fragment was purified.

a) The homeobox (180 bp) plus 131 bp of flanking sequence was amplified from pMOS-DMAHP-ABC using primers DMAHPF and DMAHPRSEH (lane 1).

b) pMOS.HD was digested with *Bam*H I and *Hind* III and the polymerase activity of the Klenow fragment used to fill the recessed termini. The 365 bp fragments were electrophoresed in a 3% (w/v) LMP agarose gel (lanes 2-5) and gel purified.

c) 5 µl of the 365 bp gel purified insert (lane 6) and 1 µl of *Sma* I digested, CIAP treated 2.9 kb pBluescript®SK(+) (lane 8) were electrophoresed in a 2% (w/v) agarose gel next to a low DNA mass marker (lane 7). It was estimated that the concentration of insert was 8 ng/µl and the concentration of vector was 200 ng/µl.

d) Six colonies of pSK(+).HD were screened for the presence of an insert by digesting the plasmid DNA with *Acc* I and electrophoresing in a 2% (w/v) gel. Five of the six colonies contained a single insert, digested into 2 fragments (lanes 9, 10, 11, 13 and 14). One colony contained multiple inserts (lane 12).

e) Two aliquots, each containing 1µg of pGEX4T3 were digested with *Eco*R I and *Not* I and electrophoresed in a 1% (w/v) LMP agarose gel (lanes 15 and 16), from which they were purified.

f) Four aliquots, each containing 200 ng of pSK(+).HD were digested with *EcoR* I and *Not* I and electrophoresed in a 2.5% (w/v) LMP agarose gel (lanes 17-20), from which 405 bp inserts were purified.

g) 5 µl of gel purified 405 bp insert (lane 23) and 1 µl of gel purified *Eco*R I/*Not* I digested 4.9 kb pGEX4T3 (lane 22) were electrophoresed in a 2% (w/v) agarose gel next to a low DNA mass marker (lane 24) and 100 ng undigested pGEX4T3 (lane 21). It was estimated that the concentration of insert was 4 ng/µl and the concentration of vector was 10 ng/µl.

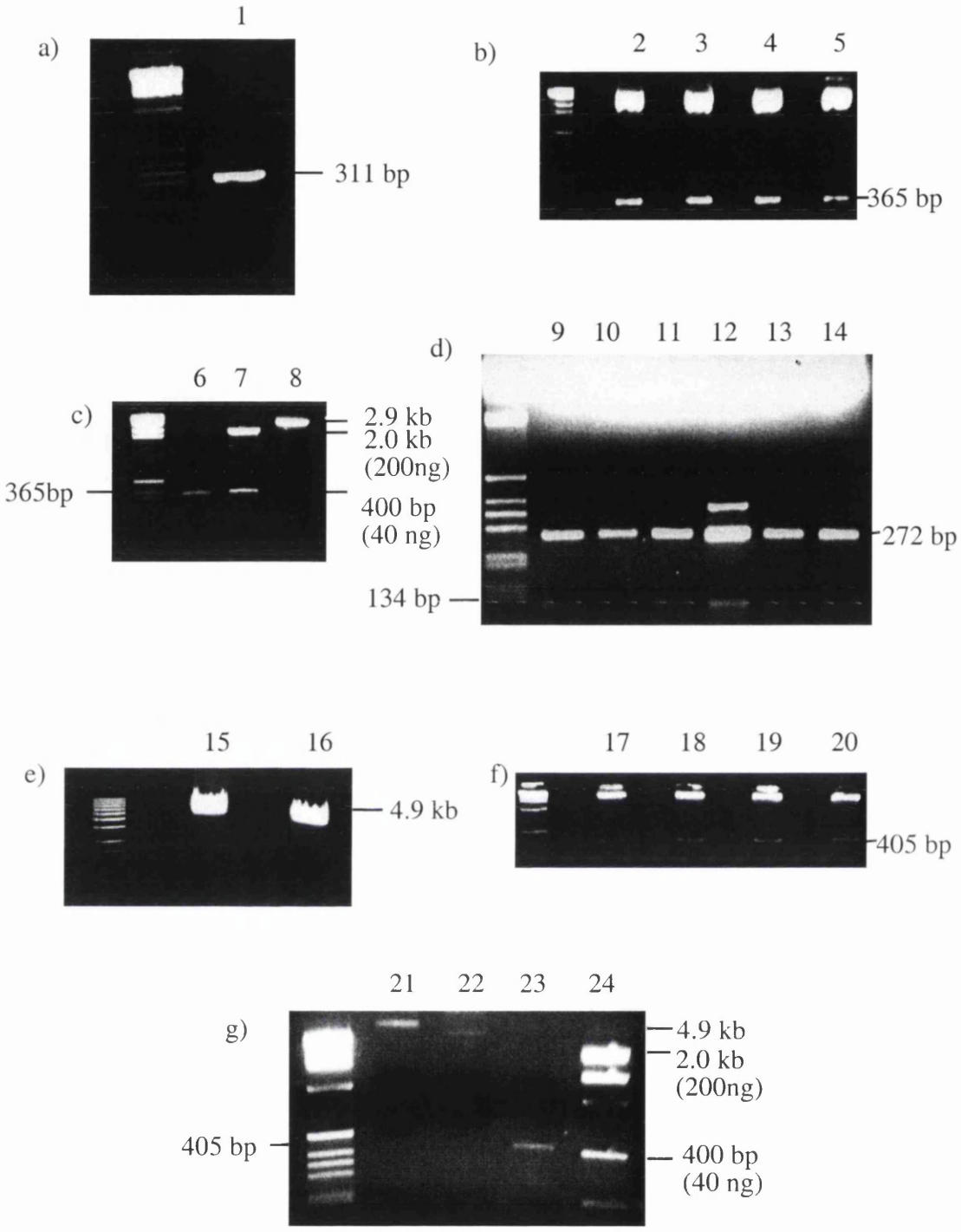

Figure 4.2 Subcloning of the homeobox into pGEX4T3

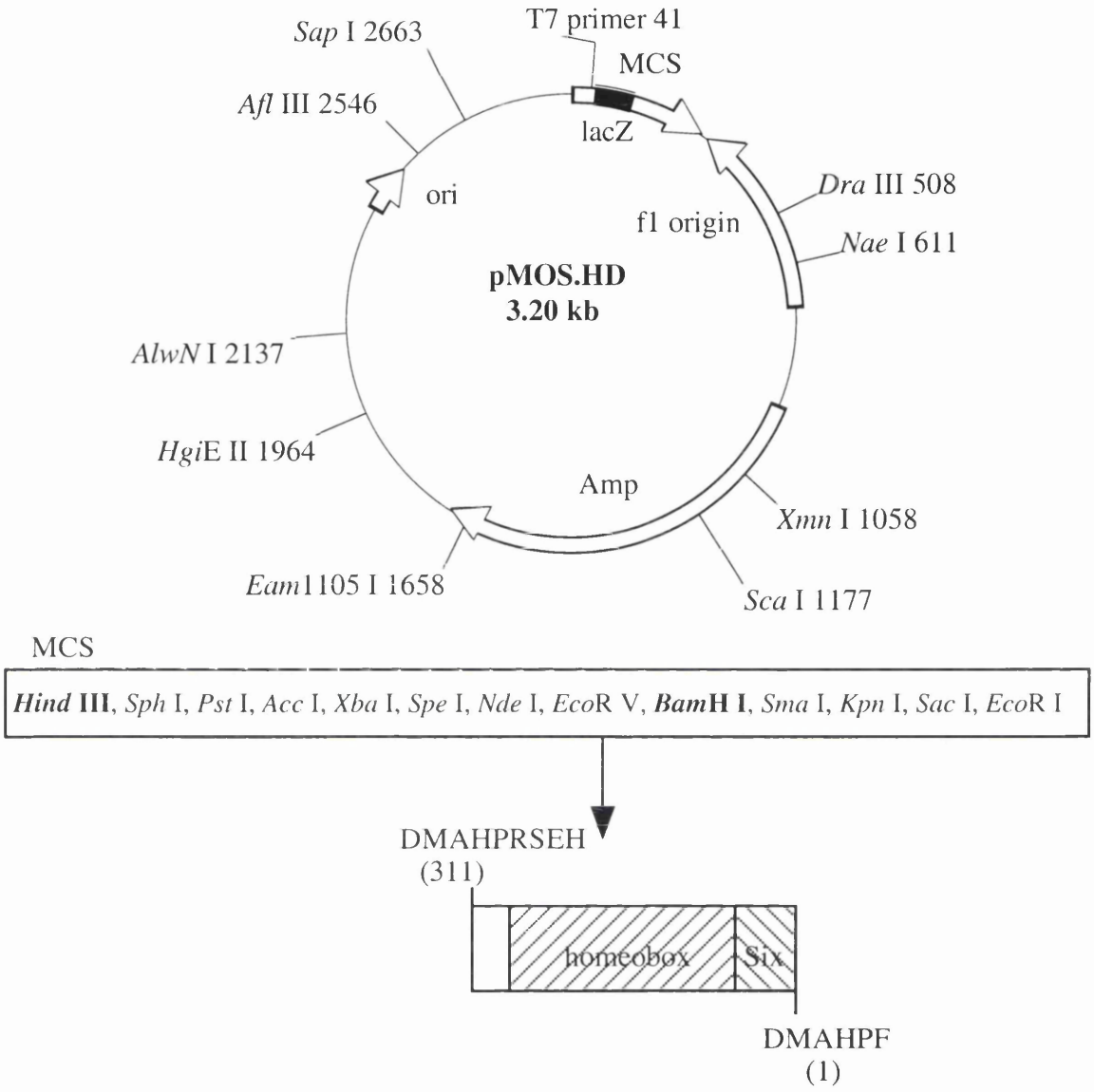

Figure 4.3 Structure of pMOS.HD

311 bp of *DMAHP*, including the entire homeobox, 52 bp of the Six box and 79 bp downstream of the homeobox were amplified from pMOS.DMAHP-ABC using primers DMAHPF and DMAHPRSEH. The PCR product was subcloned into the T-vector pMOS*Blue*. to generate pMOS.HD.

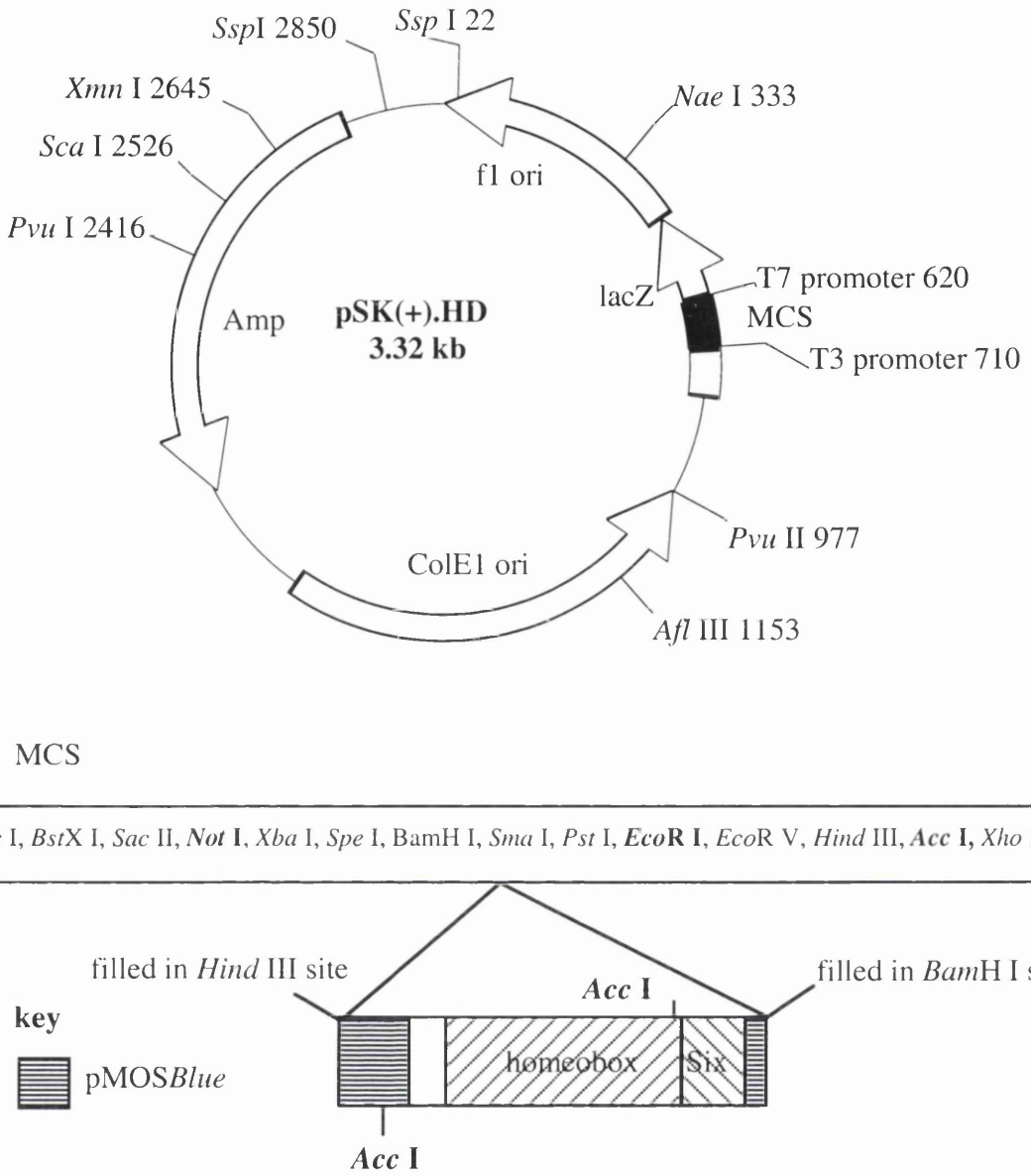

Figure 4.4 Structure of pSK(+).HD

The PCR product containing the homeobox and 52 bp of the Six box was released from pMOS.HD by digestion with *BamH* I and *Hind* III and subcloned into the *Sma* I site of pBluescript SK(+) to generate pSK(+).HD.

pGEX4T3 and pSK(+).HD were digested with *Eco*R I and *Not* I. Both, pGEX4T3 and the fragment released from pSK(+).HD were purified from 1% and 2.5% (w/v) LMP agarose gels respectively, using hot phenol (figure 4.2 e and f) and the homeobox containing fragment was ligated into pGEX4T3 to give pGEX.HD (figure 4.5). TOP10 cells were transformed with pGEX.HD. 17 out of 18 colonies analysed contained the insert. pGEX.HD was sequenced using primers pGEXF1 and pGEXB1 and this confirmed that the insert was present in the correct reading frame and no base changes had occurred (results not shown).

4.2.2 Subcloning of the Six Box into pGEX4T3

As the Six box is upstream of the RT-PCR product present in pMOS-DMAHP-ABC, it was necessary to subclone this fragment from cosmid F18894. The Six box lies within a 420 bp *Bam*H I /*Acc* I fragment. This fragment was subcloned into pBluescript® SK(+) (pSK(+).Six) to permit the use of the extensive MCS present in this vector (figure 4.6). However, as the cosmid is about 40 kb in size it was not considered possible to isolate the fragment from an agarose gel. Therefore, F18894 was digested with *Bam*H I and *Acc* I and shotgun subcloned into the *Bam*H I/*Acc* I sites of pBluescript® SK(+).

F18894 DNA was prepared using a Qiagen Maxiprep kit, following the manufacturer's instructions. 60 µg of F18894 was digested with *Bam*H I and *Acc* I and the DNA extracted using phenol chloroform and ethanol precipitated and resuspended in 10 µl H_2O. 1 µl of the DNA was electrophoresed in a 1% (w/v) agarose gel alongside 1 µl *Bam*H I /*Acc* I digested, phenol chloroform extracted, pBluescript® SK(+) (figure 4.7a). It was estimated that the concentration of pBluescript® SK(+) was 100 ng/µl and that of the 420 bp fragment released from F18894 was 9 ng/µl. Based on these estimations a ligation reaction was set up with a molar insert to vector ratio of 3:1.
DH5α cells were transformed with the construct. *Bam*H I/*Acc* I digests of 18 clones indicated that they all contained 2 inserts (results not shown). Three of the clones (clones 13, 15 and 17) contained at least one insert of about 400 bp (figure 4.7b) and were sequenced using the T7 primer (results not shown). Sequence analysis showed that the *Acc* I site of pBluescript® SK(+) had failed to cleave and that two inserts had ligated into the *Bam*H I site. Clones 15 and 17 contained the Six box fragment. This fragment was released, from clone 15, by a *Bam*H I /*Acc* I digestion (figure 4.7c) and ligated into the *Bam*H I/*Acc* I sites of pBluescript® SK(+). DH5α cells were

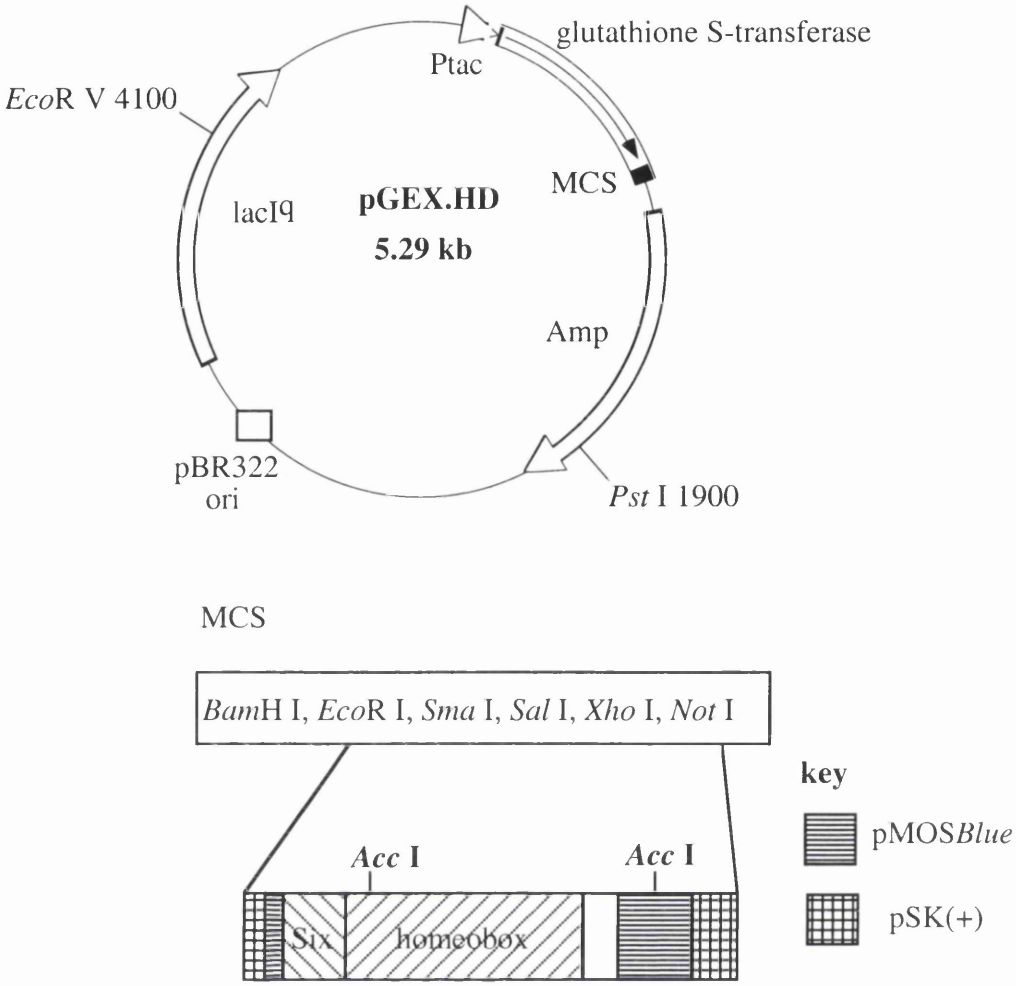

Figure 4.5 Structure of pGEX.HD

The PCR product containing the homeobox and 52 bp of the Six box was released from pSK(+).HD by digesting with *EcoR* I and *Not* I and subcloned into pGEX4T3 to generate pGEX.HD.

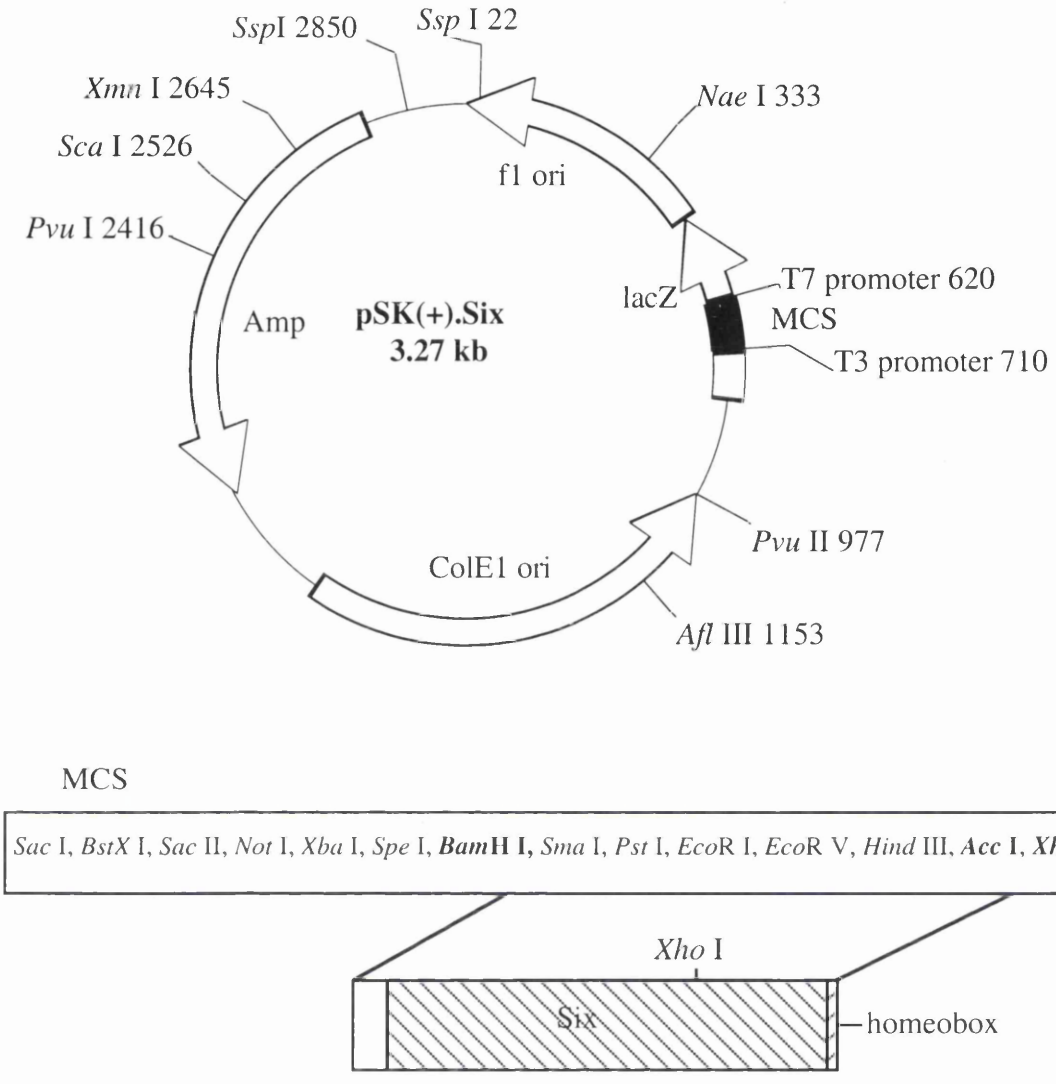

Figure 4.6 Structure of pSK(+).Six

A 420 bp *Bam*HI/*Acc* I fragment of cosmid F18894, containing the Six box, 56 bp upstream of the Six box and 17 bp of the homeobox was subcloned into pBluescriptSK(+) to generate pSK(+).Six

Figure 4.7 Subcloning of the Six box and the Six + homeobox into pGEX4T3

Vector and insert DNA used in the subcloning of the Six box and Six+homeobox into pGEX4T3 was analysed by electrophoresis in 1, 2, 2.5 or 3% (w/v) agarose gels. LMP gels were utilised when the fragment was gel purified.

a) 1 µl of *Bam*H I/*Acc* I digested F18894 DNA (lane 1) and 1 µl of gel purified *Bam*H I/*Acc* I digested pBluescript®SK(+) (lane 2) were electrophoresed in a 1% (w/v) agarose gel next to a low DNA mass marker (lane 3). It was estimated that the concentration of a 2 kb fragment of F418894 was 40 ng/µl and therefore the concentration of the 420 bp fragment was 9 ng/µl. The concentration of the vector was estimated to be 100 ng/µl.

b) Three colonies (13,15 and 17) were shown to contain one insert of about 400 bp, when digested with *Bam*H I and *Acc* I and electrophoresed in a 2.5% (w/v) agarose gel (lanes 4, 5 and 6).

c) 1 µl of gel purified *Bam*H I/*Acc* I digested clone 15, 420 bp insert (lane 8) and 1 µl of gel purified *Bam*H I/*Acc* I digested pBluescript®SK(+) (lane 7) were electrophoresed in a 1% (w/v) agarose gel next to a low DNA mass marker (lane 9). It was estimated that the concentration of the 420 bp fragment was 8 ng/µl and that of the vector was 10 ng/µl.

d) Three aliquots each containing 3 µg of pSK(+).Six were digested with *Bam*H I and partially digested with *Xho* I and electrophoresed in a 2.5% (w/v) LMP agarose gel from which the 426 bp fragment was gel purified.

e) 5 µl of gel purified 272 bp insert from *Acc* I digested pGEX.HD (lane 14) and 5 µl of gel purified *Acc* I digested pGEX.Six (lane 13) were electrophoresed in a 2% (w/v) agarose gel next to a low DNA mass marker (lane 15). It was estimated that the concentration of the insert was 2 ng/µl and that of the vector was 40 ng/µl.

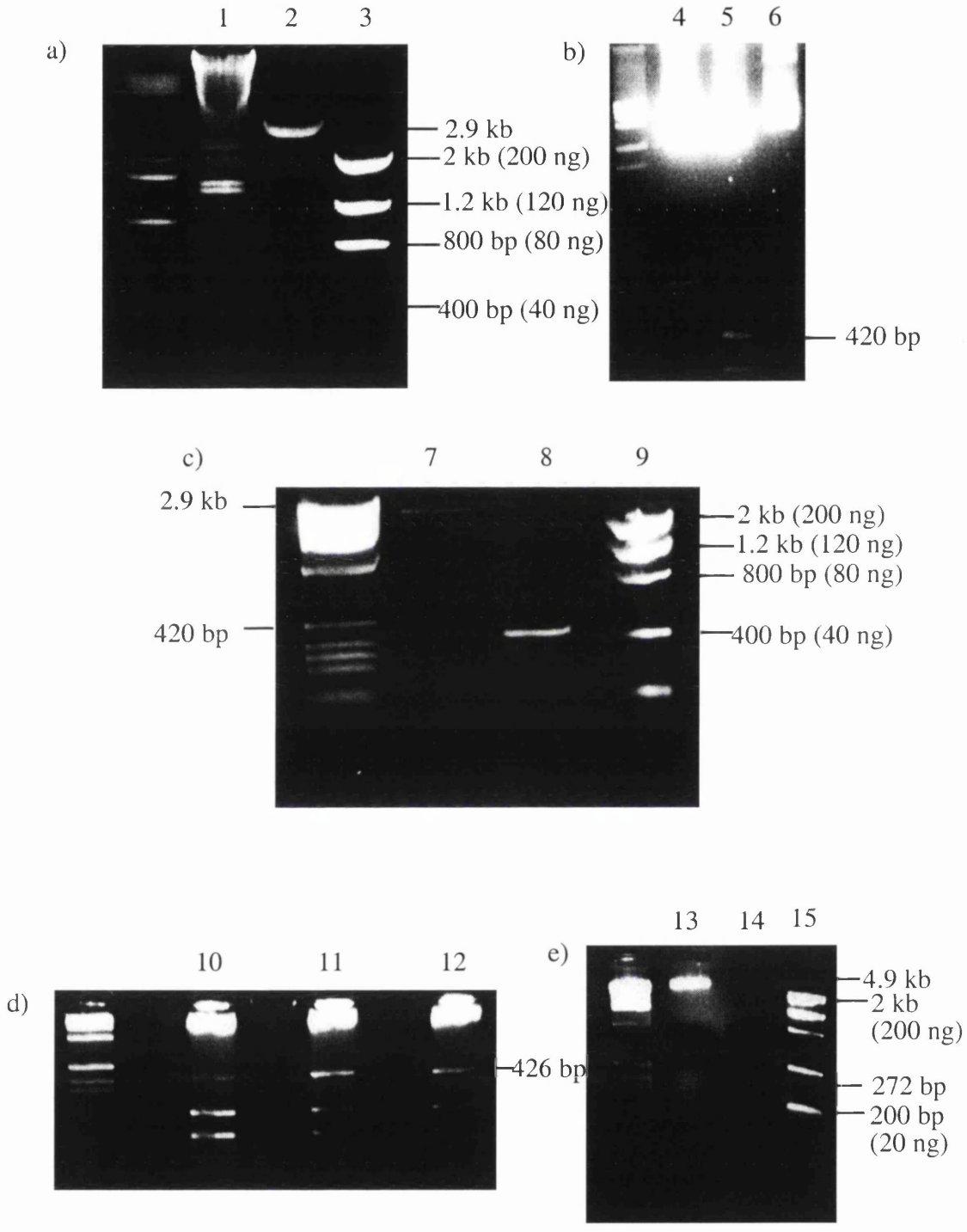

Figure 4.7 Subcloning of the Six box and the Six+homeobox into pGEX4T3

transformed with pSK(+).Six. Fortunately, this time, the *Acc* I site was cleaved and 13 out of the 17 clones tested contained the correct sized insert. pSK(+).Six was sequenced with T7 and T3 primers and the expected sequence of pSK(+).Six was confirmed (results not shown).

pGEX4T3 and pSK(+).Six were digested with *Bam*H I and *Xho* I. Although an *Xho* I site exists within the Six box (figure 4.6), *Xho* I cuts with only 70-100% efficiency in the restriction endonuclease buffer used (Boehringer Mannheim buffer B). Therefore, due to incomplete digestion by *Xho* I, three different sized fragments were released from pSK(+).Six, the largest (426 bp) being the one required (figure 4.7d). Both, pGEX4T3 and the 426 bp fragment released from pSK(+).Six were purified from 1% and 2.5% (w/v) LMP agarose gels respectively, using hot phenol and the Six box containing fragment was ligated into pGEX4T3 to give pGEX.Six (figure 4.8). TOP10 cells were transformed with pGEX.Six. 15 out of 18 colonies analysed contained the insert. pGEX.Six was sequenced using primers pGEXF1 and pGEXB1 and this confirmed that the insert was present in the correct reading frame and no base changes had occurred (results not shown).

4.2.3 Subcloning of the Six+HD into pGEX4T3

pGEX.Six (figure 4.8) was digested with *Acc* I, to linearise the construct, and CIAP treated to prevent it religating and the DNA extracted using phenol/chloroform. pGEX.HD (figure 4.5) was also digested with *Acc* I which released an insert containing most of the homeobox plus 80 bp of 3' flanking sequence, which included 53 bp of exon B and pMOS*Blue* sequence from the T cloning site to the *Acc* I site. This insert was purified from a 3% (w/v) LMP agarose gel and ligated into pGEX.Six to give pGEX.Six+HD (figures 4.7e and 4.9). TOP10 cells were transformed with the construct. Clones were tested for the presence of an insert by digesting the construct with *Sac* II as a *Sac* II site exists in both the Six box and in the homeobox (figure 4.9). One of the positive clones was sequenced using primers pGEXF1 and pGEXB1 and this confirmed that the insert was present in the correct reading frame and no base changes had occurred (results not shown).

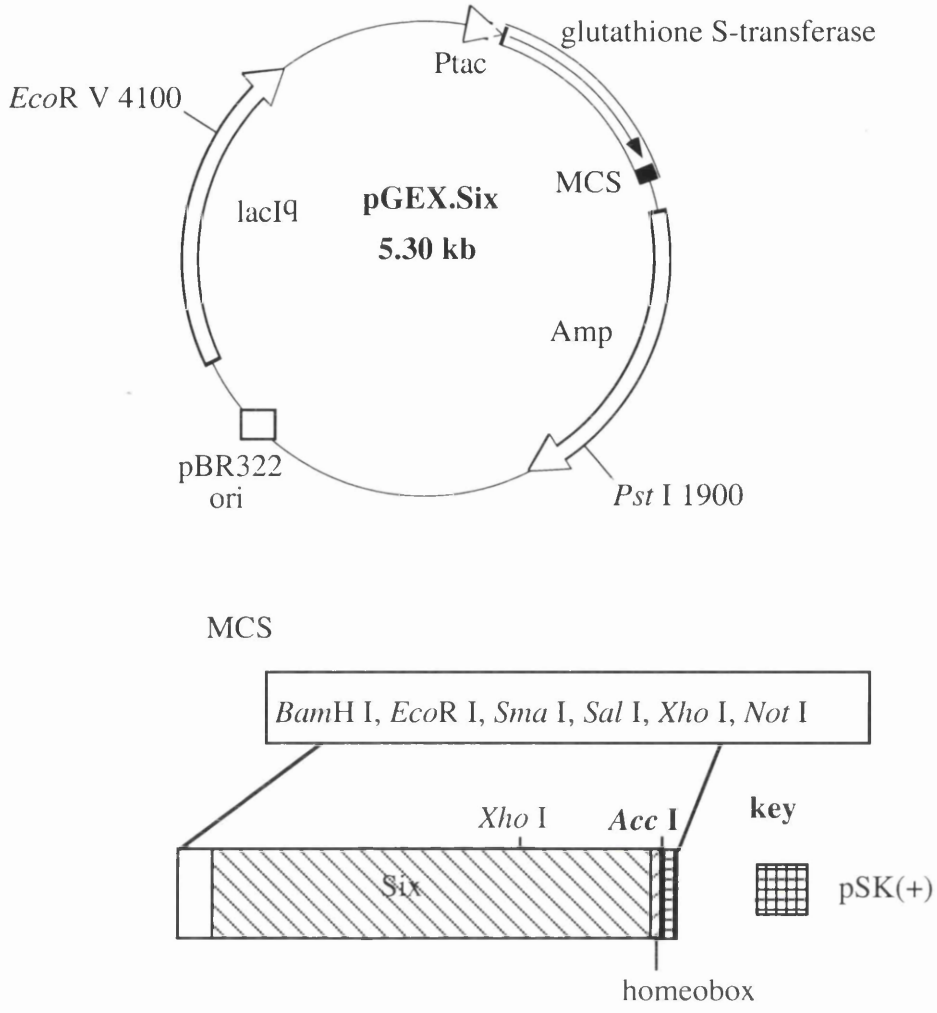

Figure 4.8 Structure of pGEX.Six

A 426 bp fragment, containing the Six box, 56 bp of upstream DNA and 17 bp of the homeobox was released from pSK(+)Six by digestion with *Bam*H I and *Xho* I and subcloned into the *Bam*H I/*Xho* I sites of pGEX4T3 to generate pGEX.Six..

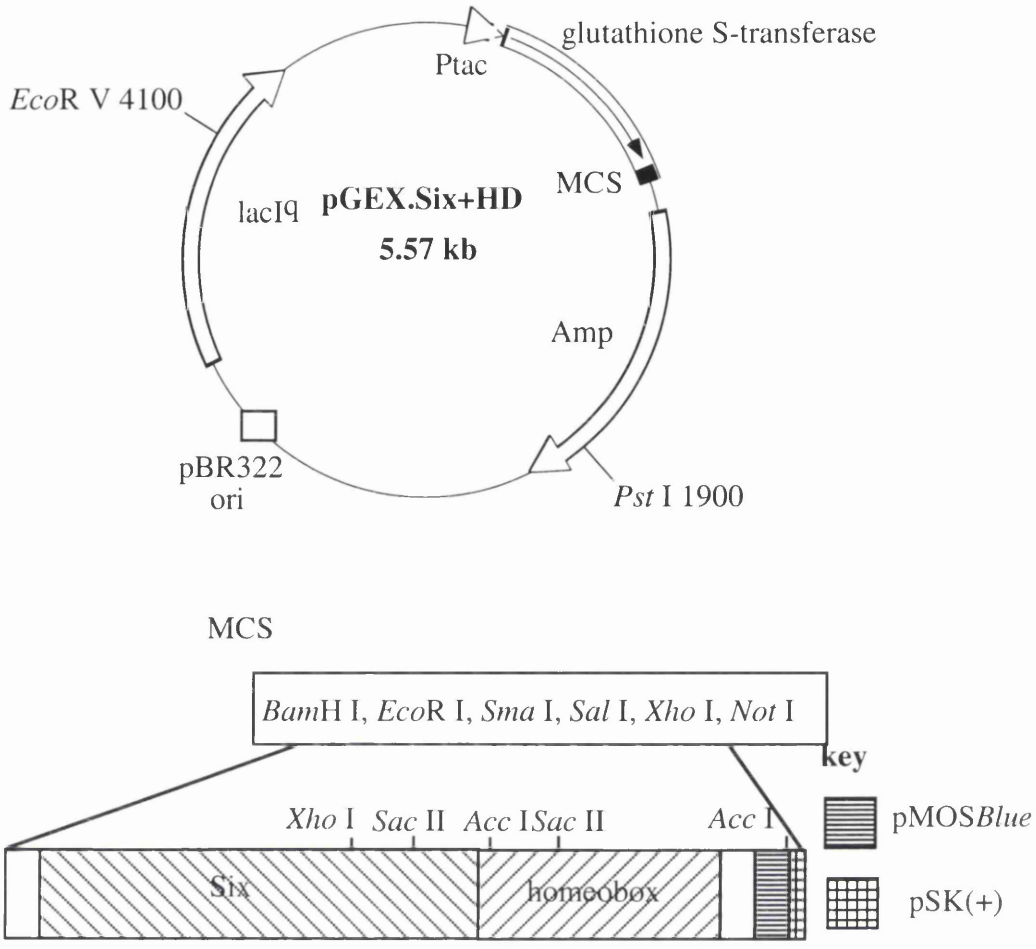

Figure 4.9 Structure of pGEX.Six+HD

The homeobox was released from pGEX.HD by digestion with *Acc* I and subcloned into the *Acc* I site of pGEX.Six to generate pGEX.Six+HD which contains the entire Six and homeobox plus 136 bp of flanking DNA.

4.3 Expression of Recombinant GST-DMAHP Proteins

Recombinant GST-DMAHP proteins were expressed from the pGEX4T3 constructs described above (pGEX.HD, pGEX.Six and pGEX.Six+HD). GST protein tag was also expressed both as a positive control for the expression experiments and also as a negative control for the binding experiments (chapter 5). The expected molecular weights of the recombinant proteins were estimated using MacVector: GST-HD (41 kDa), GST-Six (45 kDa) and GST-Six+HD (53 kDa).

4.3.1 Optimisation of the Expression of Recombinant GST-DMAHP Proteins

A protein expression time course (section 2.3.12.2) was required to determine the optimum growth and recombinant protein synthesis conditions for each protein. 1 mM IPTG was used to induce expression of recombinant protein for 5 hours at 37°C and readings were taken every hour with a sixth reading, after overnight expression (21 hours). SDS-PAGE analysis of the crude cell lysate was used to detect overexpressed recombinant proteins. The growth kinetics of each recombinant GST fusion protein were analysed by measuring the optical density of the cells at each time point.

a) GST -HD protein expression time course

No overexpressed GST-HD protein was detected by SDS-PAGE analysis of the crude cell lysate, produced in the time course experiment (results not shown), although GST protein of the expected size (29 kDa) was clearly visible in the supernatant fraction of the pGEX4T3 cultures (figure 4.10a). Induction of the expression of GST protein had little effect on the growth of the culture (figure 4.11). However, the growth rate of the culture of TOP10 cells transformed with pGEX.HD and induced with 1 mM IPTG was reduced, compared to the non-induced culture, which grew at about the same rate as both GST cultures.

In an attempt to produce detectable amounts of GST-HD, cultures were induced for longer periods of time (up to 30 hours), with 0.5 mM, 1 mM, 2 mM and 5 mM concentrations of IPTG. No overexpressed GST-HD protein was detected by SDS-PAGE analysis of the crude cell lysate, in any of the samples (results not shown).

It was hypothesised that the GST-HD protein was toxic to the cells, which would explain the reduced growth rate of the cultures induced with IPTG. It is known that basal level expression occurs when using the *tac* inducible promoter present in

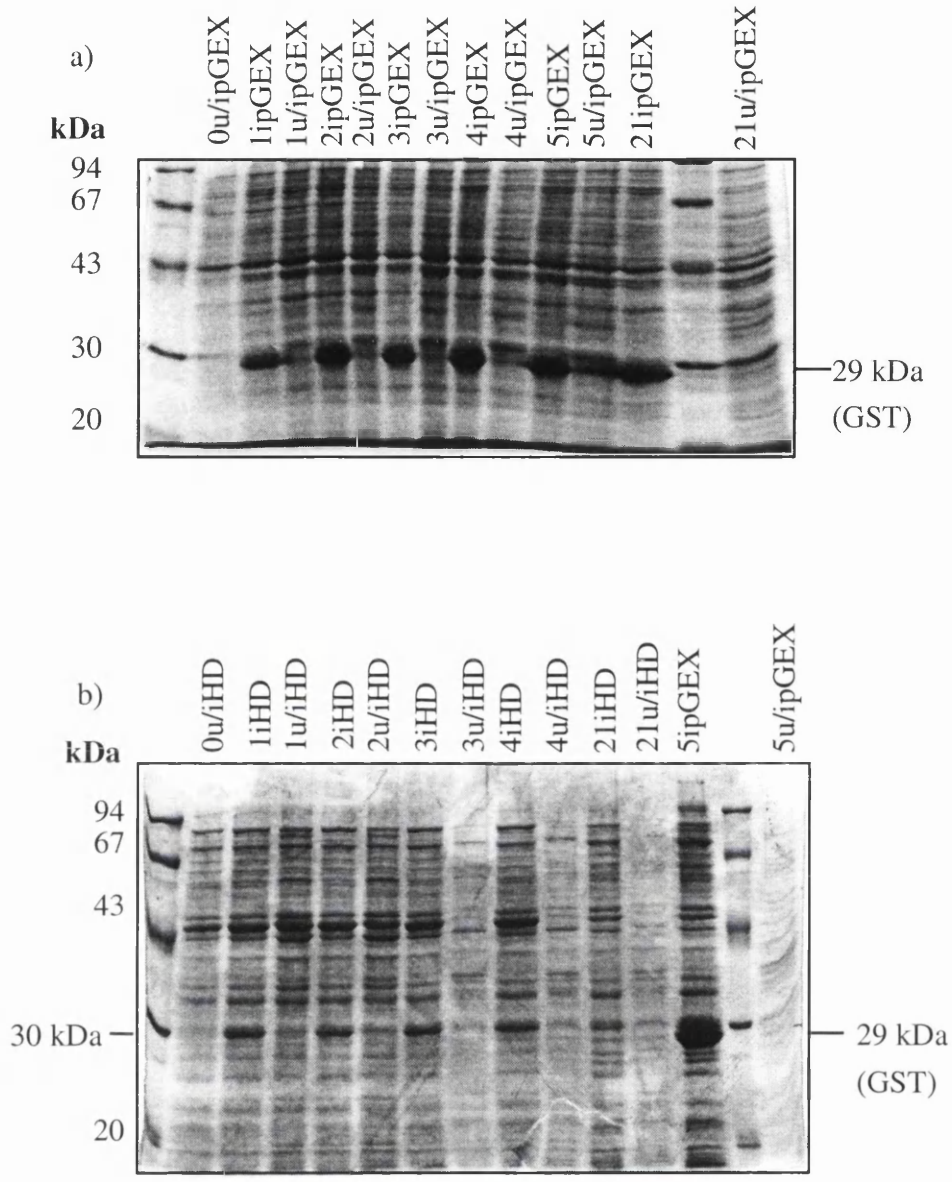

Figure 4.10 SDS-PAGE analysis of the protein extracted from the supernatant fraction of GST and GST-HD protein expression time course experiments.

6% (v/v) of the supernatant fraction of 1 ml aliquots from 100 ml cultures of pGEX4T3 and pGEX.HD were analysed by SDS-PAGE analysis in a 10% (w/v) SDS polyacrylamide gel. GST (pGEX) expression was induced in the absence of glucose and GST-HD (HD) was induced in the presence of 2% (w/v) glucose. 0-21 = aliquot taken 0-21 hours after addition of IPTG, i = 1 mM IPTG added, u/i = no IPTG added. 29 kDa GST protein can be seen in lanes: 1ipGEX, 2ipGEX, 3ipGEX, 4ipGEX, 5ipGEX and 6ipGEX, but not in any of the u/i lanes. A 30 kDa protein can be seen in lanes: 1iHD, 2iHD, 3iHD, 4iHD and 5iHD, decreasing in concentration with time. It is not visible in any of the u/i lanes.

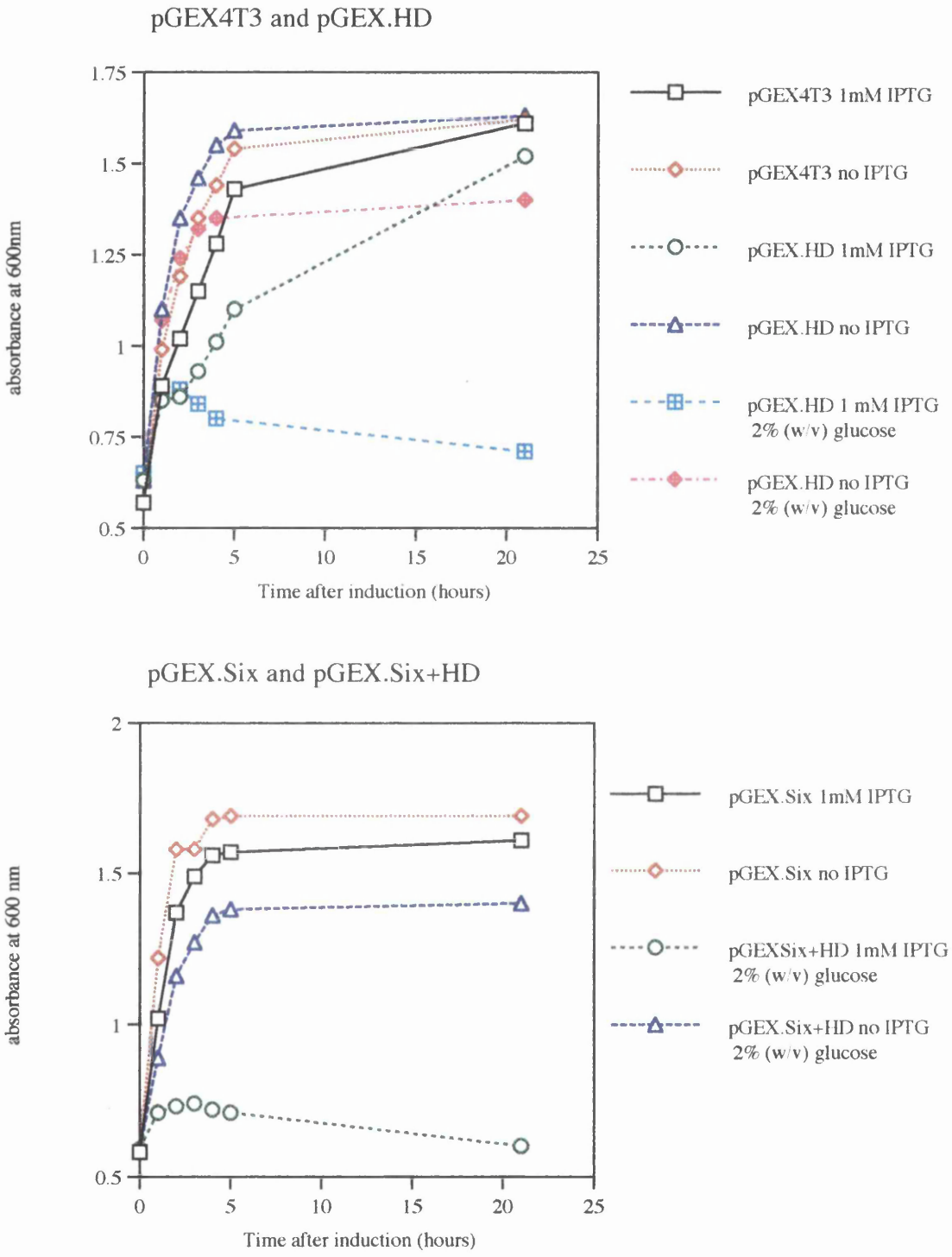

Figure 4.11 Growth curves of pGEX and pGEX-DMAHP cultures

The optical density of 1ml of cells at 600 nm was plotted against time.

pGEX4T3 and may be due to the presence of a *lac* promoter located upstream (Pharmacia Biotech). The cells may already have been dying when the IPTG was added and maximum protein production may have taken place. The addition of 2% (w/v) glucose decreases the basal level expression associated with the *lac* promoter (Pharmacia Biotech). Therefore, a protein expression time course was performed in the presence of 2% (w/v) glucose. SDS-PAGE analysis did not detect an overexpressed protein of the expected size (41 kDa). However a 30 kDa protein (which is just heavier than the GST protein, alone) was detected in the supernatant fraction of the crude cell lysate (figure 4.10b). As a truncated GST-DMAHP protein had previously been reported (Winchester, 1997), it was decided to investigate this protein by affinity purification (section 4.3.2b) and western blot analysis (section 4.3.3). The growth rate of the culture of TOP10 cells transformed with pGEX.HD induced with 1 mM IPTG and 2% (w/v) glucose was reduced, with respect to uninduced culture and the absorbance at 600 nm actually decreased after 4 hours induction (figure 4.11). The amount of protein produced, as visualised by SDS-PAGE analysis also decreased with time (figure 4.10b).

b) GST-Six protein expression time course

The growth rate of the culture expressing GST-Six was reduced only slightly, with respect to the uninduced culture (figure 4.11). SDS-PAGE analysis detected a protein of 42 kDa in the pellet fraction of the crude cell lysate (figure 4.12). The predicted size was 45 kDa. From two hours post-induction, GST-Six could also be visualised in the pellet fractions containing no IPTG, due to basal level expression (section 4.3.1a).

c) GST-Six+HD expression time course

2% (w/v) glucose was added to the culture to prevent basal level expression, which had proven to be a problem when expressing GST-HD protein. No overexpressed GST-Six+HD protein was detected by SDS-PAGE analysis of the crude cell lysate (results not shown). The growth rate of the culture of TOP10 cells transformed with pGEX.Six induced with 1 mM IPTG and 2% (w/v) glucose was reduced, with respect to the uninduced culture and the GST cultures, with the absorbance at 600 nm decreasing after 4 hours of induction (figure 4.11). This suggested that any expressed recombinant GST fusion protein was toxic to the cells. As no overexpressed protein could be detected by SDS-PAGE analysis, even with the addition of 2% (w/v) glucose, the presence of recombinant GST-Six+HD was investigated by western blot analysis (section 2.3.16). GST antiserum detected a doublet of 52 and 50 kDa and a

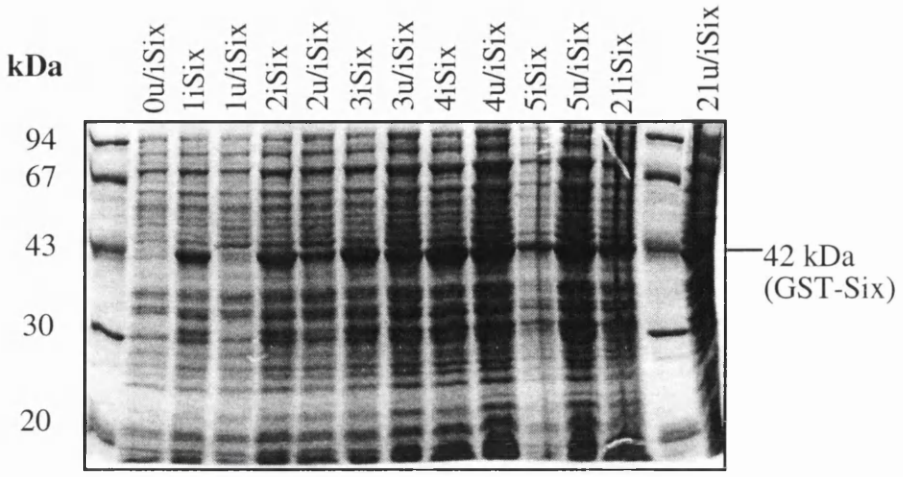

Figure 4.12 SDS-PAGE analysis of the protein extracted from the pellet fraction of a GST-Six protein expression time course

5% (v/v) of the pellet fraction of 1 ml aliquots from a 100 ml culture of pGEX.Six were analysed by SDS-PAGE analysis in a 10% (w/v) gel. 0-21 = aliquots taken 0-21 hours after the addition of IPTG. i = 1mM IPTG added, u/i = no IPTG added. The 42 kDa GST-Six (Six) recombinant protein can be seen in lanes: 1iSix, 2iSix, 3iSix,4u/iSix, 4iSix, 5u/iSix, 5iSix, 21u/iSix and 21iSix.

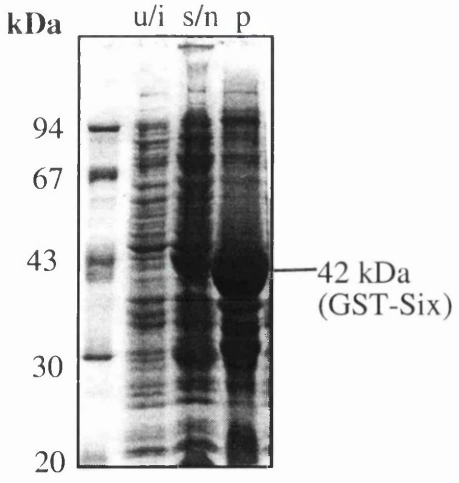

Figure 4.13 SDS-PAGE analysis of the protein extracted from a large scale culture of GST-Six protein

0.02% (v/v) of a 500 ml uninduced (u/i) culture of pGEX.Six, 0.10% (v/v) of the supernatant (s/n) fraction and 0.135% (v/v) of the pellet (p) fraction of a culture induced for 4 hours with 1mM IPTG at 30°C were electrophoresed in a 10% (w/v) SDS polyacrylamide gel. The 42 kDa GST-Six protein is visible in the pellet fraction.

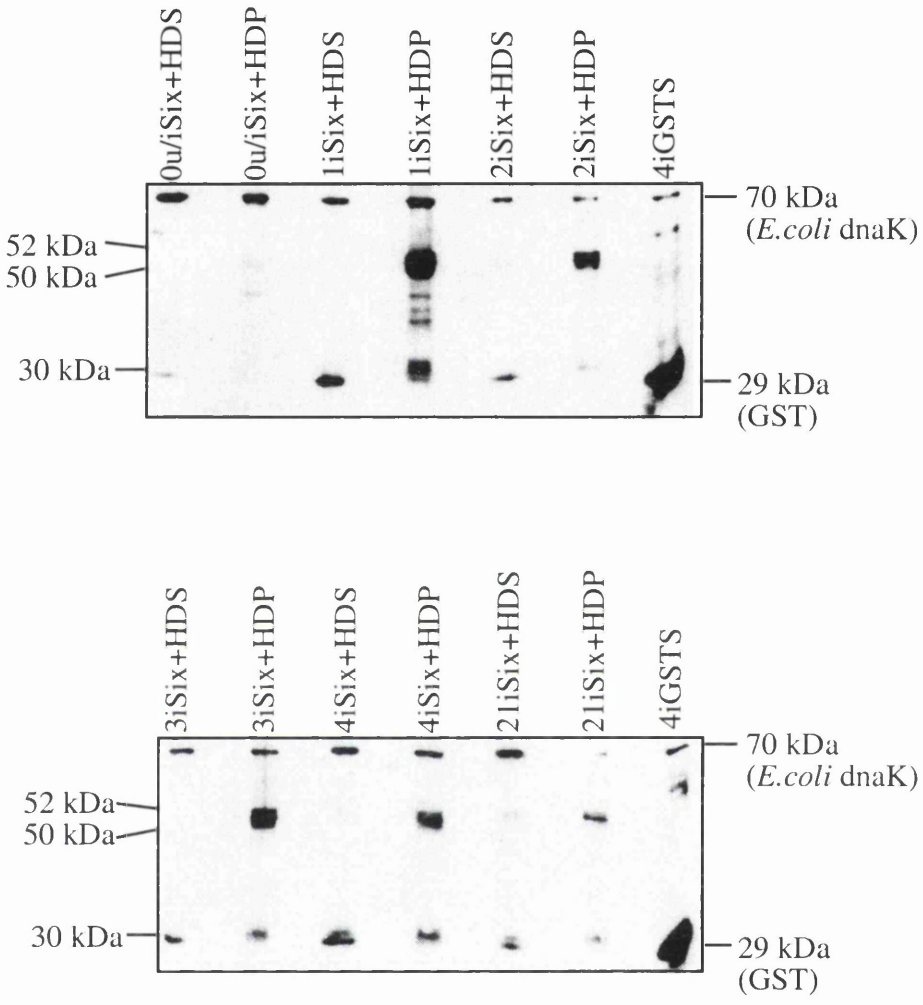

Figure 4.14 Western blot analysis of the protein extracted from GST-Six+HD protein expression time course

5% (v/v) of the pellet (P) fraction and 6% (v/v) of the supernatant (S) fraction of 1 ml aliquots from 100 ml cultures of pGEX-Six+HD (Six+HD) and pGEX (GST) were electrophoresed in a 10% (w/v) SDS polyacrylamide gel and electroblotted onto Hybond-N ECL nitrocellulose membrane. The membrane was probed with GST antiserum 0-21 = aliquot taken 0-21 hours after the addition of IPTG. i = 1 mM IPTG added, u/i = no IPTG added. The GST-Six+HD protein is visible in all pellet lanes, decreasing in concentration,with time.

30 kDa protein in the pellet fractions of the crude cell lysates and the concentration of these proteins decreased with time (figure 4.14). The expected size of the protein was 53 kDa. A 70 kDa protein was also detected, in both supernatant and pellet fractions. The 30 kDa protein is presumably a partial GST-Six+HD protein and the 70 kDa is thought to be a product of the *E.coli dnaK* gene, which is involved in the degradation of abnormal proteins in *E.coli* (Pharmacia Biotech). The supernatant fractions also contained a 29 kDa protein, which was assumed to be GST.

4.3.2 Large Scale Expression and Purification of Recombinant GST-DMAHP Proteins

To produce greater amounts of recombinant GST fusion protein, for affinity purification, 500 ml cultures were grown (section 2.3.12.3). Sonication was used to lyse the cells, rather than freeze/thawing as this had previously been shown to be a more efficient method, when using larger cultures (Winchester, 1997). Recombinant GST fusion protein was affinity purified using Glutathione Sepharose® 4B beads (section 2.3.12.5). 20 ml of sonicated cell lysate or 8 ml of solubilised pellet fraction was mixed with 1.6 ml of 50% (w/v) Glutathione Sepharose® 4B beads resuspended in MTPBS buffer. Recombinant GST fusion protein was eluted from the glutathione Sepharose® 4B beads by the addition of 6 ml of elution buffer [50 mM Tris (pH 8.0), 120 mM NaCl, 10 mM DTT, 1 mM PMSF, 20 mM reduced glutathione].

a) GST expression and purification

From the growth kinetics experiments carried out when optimising the expression of the 3 fusion proteins, 4 hours was chosen as the optimum expression time for GST protein (figure 4.10a). SDS-PAGE analysis of affinity purified protein detected a 29 kDa protein (figure 4.15a).

b) GST-HD expression and purification

As the protein had been shown to be toxic to the cells, cultures were induced at 30°C for 1 hour. SDS-PAGE analysis of affinity purified protein detected a doublet of 38 kDa and 40 kDa (figure 4.15b). The expected size of the protein was 41 kDa.

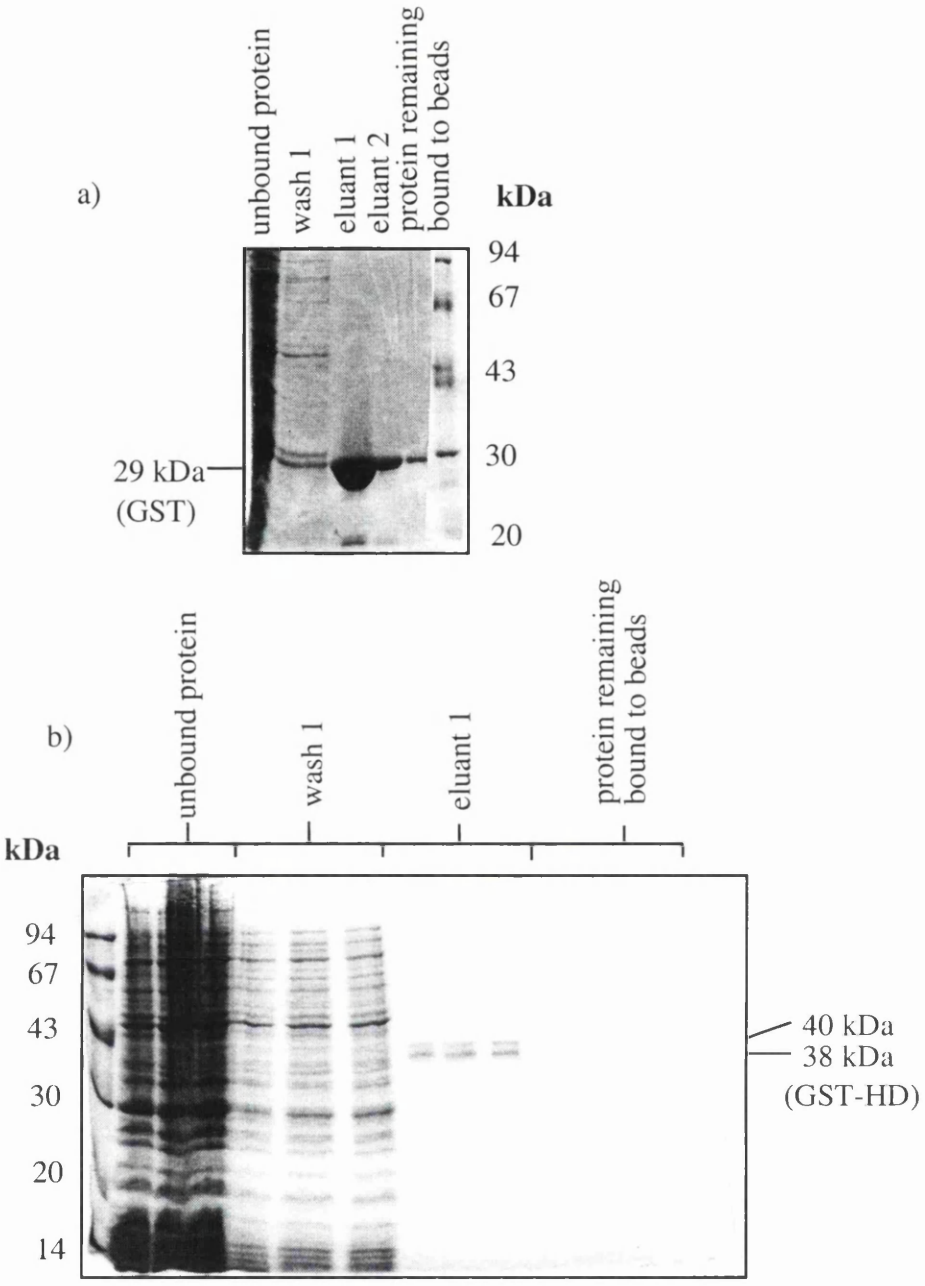

Figure 4.15 SDS-PAGE analysis of the stages of affinity purification of GST and GST-HD

0.04% (v/v) unbound protein, 0.07% (v/v) wash 1, 0.13% (v/v) eluted protein and 0.13% (v/v) protein remaining bound to the beads were analysed by SDS-PAGE analysis in a 10% (w/v) gel. To obtain sufficient protein for use in gel retardation assays, three 400 ml cultures of pGEX.HD were prepared.

c) GST-Six expression and purification

As GST-Six was present in the pellet fraction of the cell lysate during the time course experiment, cells were induced at 30°C, as lower growth temperatures have been shown to increase fusion protein solubility (Schein, 1989). Four hours was chosen as the optimum time for protein expression and 2% (w/v) glucose was added to the culture to reduce basal level expression. SDS-PAGE analysis showed that GST-Six was still present in the pellet fraction (figure 4.13) and therefore inclusion bodies were solubilised (section 2.3.12.4). The stages of inclusion body solubilisation were analysed by SDS-PAGE analysis (figure 4.16) and the 42 kDa protein could be visualised in the solubilised pellet fraction. SDS-PAGE analysis of affinity purified GST-Six detected a major band of 42 kDa with a fainter band running at about 41 kDa. A faint band of 30 kDa, which is presumably GST protein plus a small amount of DMAHP, that has separated from the main fusion protein by proteolytic cleavage, could also be seen (figure 4.17). Much of the protein remained on the Glutathione Sepharose® 4B beads.

d) GST-Six+HD expression and purification

Cells were induced at 30°C, to increase the solubility of the protein. One hour was chosen as the optimum time for protein expression and 2% (w/v) glucose was added to the culture to reduce basal level expression. No recombinant GST-Six+HD was detected by SDS-PAGE analysis (result not shown). However, western blot analysis, with R254 DMAHP antiserum, to confirm that the overexpressed protein was a DMAHP protein, showed that the recombinant protein was now in the supernatant fraction (figure 4.18) which would facilitate affinity purification. SDS-PAGE of affinity purified protein detected a doublet of 52 kDa and 50 kDa (figure 4.19).

4.3.3 Western Blot Analysis of GST-HD and GST-Six Proteins

To confirm that the proteins visualised by SDS PAGE analysis were GST-DMAHP recombinant proteins they were transferred to Hybond™-N ECL nitrocellulose membrane and analysed with polyclonal antisera (section 2.3.16). GST antiserum and R254 DMAHP antiserum were used as primary antisera, with HRP-conjugated secondary antibodies, to detect the recombinant GST-DMAHP proteins.

Affinity purified GST-Six, GST-HD and GST protein were analysed. GST antiserum detected the 42 kDa and the 30 kDa affinity purified GST-Six proteins, the affinity

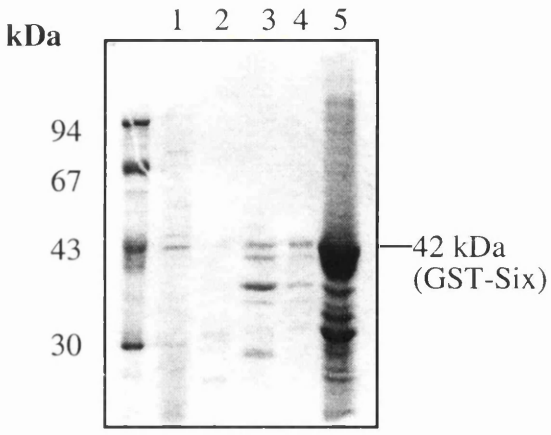

Figure 4.16 SDS-PAGE analysis of GST-Six obtained from inclusion bodies

GST-Six protein was obtained by solubilising the inclusion bodies present in the pellet fraction of the cell lysate. Aliquots of the supernatants produced at each stage of the solubilisation procedure were electrophoresed in a 10% (w/v) SDS polyacrylamide gel. The lanes are as follows: 0.02% (v/v) of supernatant 1 (lane 1), 0.02% (v/v) of supernatant 2 (lane 2), 0.07% (v/v) of supernatant 3 (lane 3), 0.07% (v/v) of supernatant 4 (lane 4) and 0.1% (v/v) of solubilised and dialysed GST-Six (lane 5).

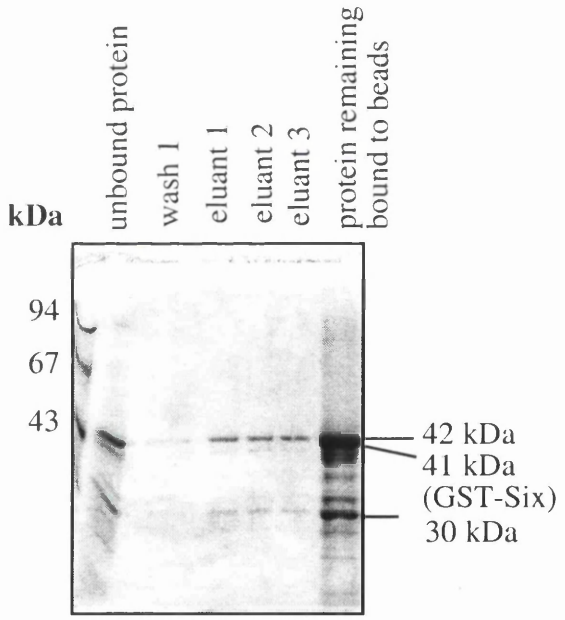

Figure 4.17 SDS-PAGE analysis of the stages of affinity purification of GST-Six

0.04% (v/v) unbound protein, 0.07% (v/v) wash 1, 0.13% (v/v) eluted protein and 0.13% (v/v) protein remaining bound to the beads were analysed by SDS-PAGE analysis in a 10% (w/v) gel. Despite 3 elutions the majority of the protein remained bound to the beads.

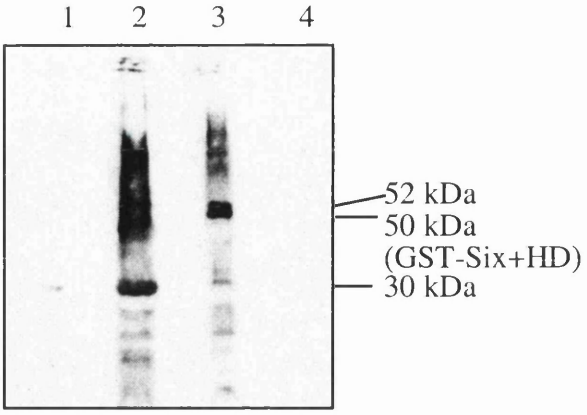

Figure 4.18 Western blot analysis of the protein extracted from a large scale culture of GST-Six+HD protein

0.01% (v/v) of an uninduced 500 ml culture of pGEX-Six+HD (lane 1), 0.05% (v/v) of the pellet fraction (lane 2) and 0.1% (v/v) of the supernatant fraction (lane 3) of a culture overexpressing GST-Six+HD and 0.1% (v/v) of the supernatant fraction of a culture overexpressing GST (lane 4) were analysed by western blotting and probing with R254 DMAHP antiserum. GST-Six+HD was detected in the supernatant fraction (lane 3).

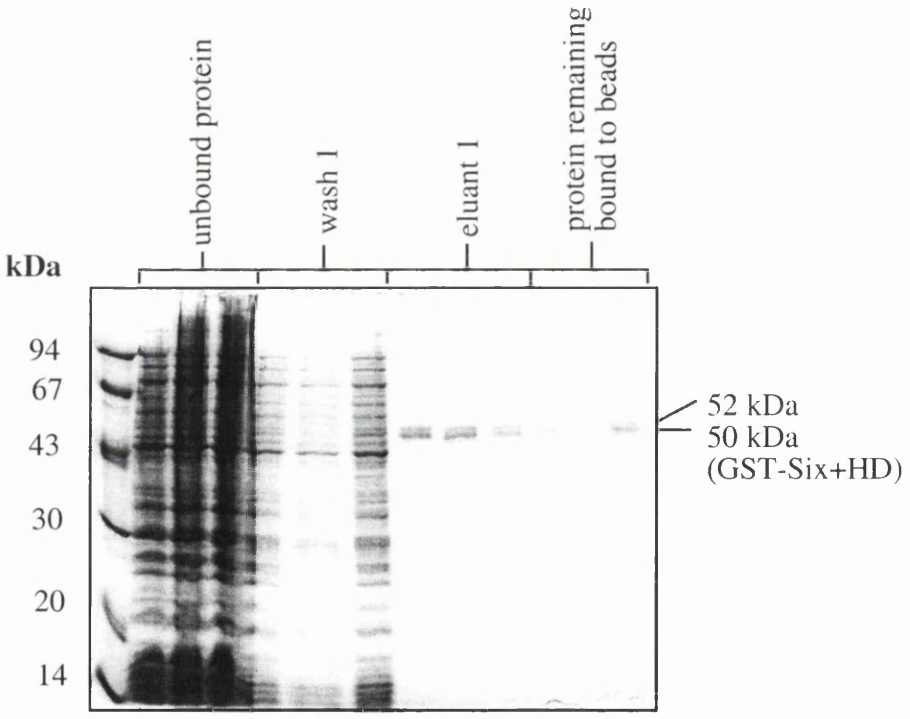

Figure 4.19 SDS-PAGE analysis of the stages of affinity purification of GST and GST-Six+HD

0.04% (v/v) unbound protein, 0.07% (v/v) wash 1, 0.13% (v/v) eluted protein and 0.13% (v/v) protein remaining bound to the beads were analysed by SDS-PAGE analysis in a 10% (w/v) gel. To obtain sufficient protein for use in gel retardation assays, three 400 ml cultures of pGEX.HD were prepared.

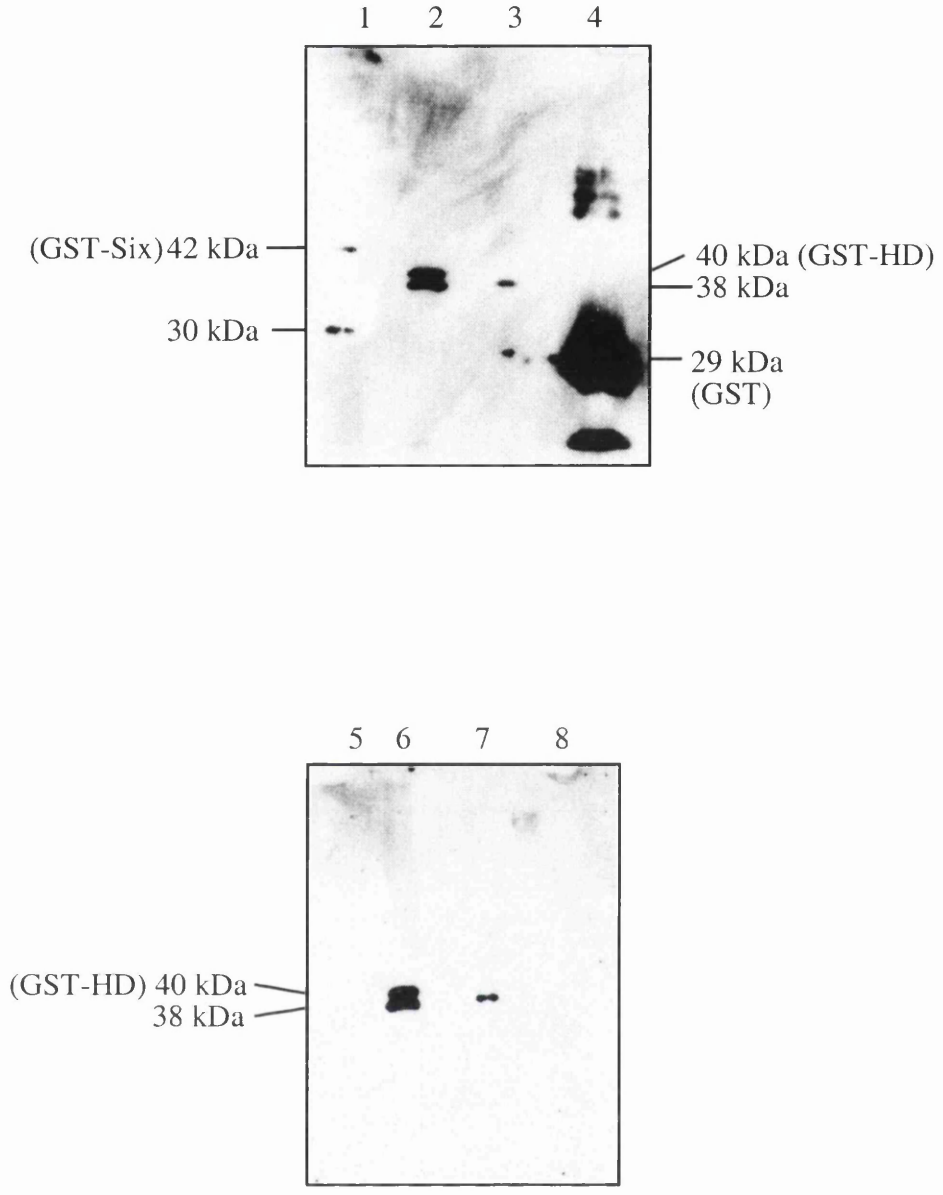

Figure 4.20 Western blot analysis of affinity purified GST-Six, GST-HD and GST recombinant proteins

250 ng of affinity purified GST-Six (lanes 1and 5), 200 ng of affinity purified GST-HD (lanes 2 and 6) and 1000 ng of affinity purified GST (lanes 4 and 8) were electrophoresed in a 10% (w/v) SDS polyacrylamide gel and electroblotted onto Hybond-N ECL nitrocellulose membrane. Lanes 1-4 were probed with GST antiserum and lanes 5-8 with R254 DMAHP antiserum. Lanes 3 and 7 are not applicable to this experiment.

purified GST-HD protein doublet and affinity purified GST protein (figure 4.20). A 30 kDa protein in the GST-HD lane could also be seen after a 10 minute exposure (result not shown). R254 DMAHP antiserum detected only the affinity purified homeodomain doublet (figure 4.20). As R254 DMAHP antiserum was raised against only a small part of the Six domain it was not expected to detect GST-Six.

4.4 Concentrating Recombinant GST-DMAHP Proteins

4.4.1 Estimating the Concentration of Recombinant GST-DMAHP Proteins

The concentrations of affinity purified GST-HD and GST-Six were estimated from figure 4.21 as described in section 2.3.12.6. 8 µl affinity purified protein was electrophoresed in a 10% (w/v) SDS-polyacrylamide gel, alongside Coomassie Blue-stained BSA standards. It was estimated that the concentration of GST-HD was 25 ng/µl and GST-Six was 31 ng/µl.

4.4.2 Concentrating the Affinity Purified Recombinant Proteins

Although the affinity purified proteins produced were relatively pure, they were not very concentrated. To enable an intensive study of their functional properties it was necessary to obtain them at a higher concentration.

a) 2 ml of affinity purified GST-HD and GST-Six were reduced to 200 µl using a Centricon 10 membrane unit, according to the manufacturer's instructions. 8 µl concentrated protein was electrophoresed in a 10% (w/v) SDS-polyacrylamide gel, alongside Coomassie Blue-stained BSA standards (figure 4.21). The concentration of GST-HD was estimated to be 2.5 ng/µl. No GST-Six recombinant protein was visible on the gel.

b) 500 µl cold 100% (v/v) ethanol and acetone were used to precipitate affinity purified GST-HD, GST-Six and GST-Six+HD (section 2.3.12.7b). The pellet was dried and resuspended in 50 µl cold 50 mM Tris (pH 8.0), 120 mM NaCl, 10 mM DTT, 1 mM PMSF and 10.5% (v/v) glycerol. The sample was either snap frozen immediately or dialysed against 50 mM Tris (pH 8.0), 120 mM NaCl and 10.5% (v/v) glycerol, with 10 mM DTT and 1 mM PMSF being added before being snap frozen. Dialysis was performed to ensure that the correct salt concentration was maintained, for further experiments. 10 µl aliquots of each of the 4 samples and unconcentrated GST-HD were analysed by SDS-PAGE. The two samples that had been precipitated but not dialysed contained more protein than the unconcentrated sample. The sample

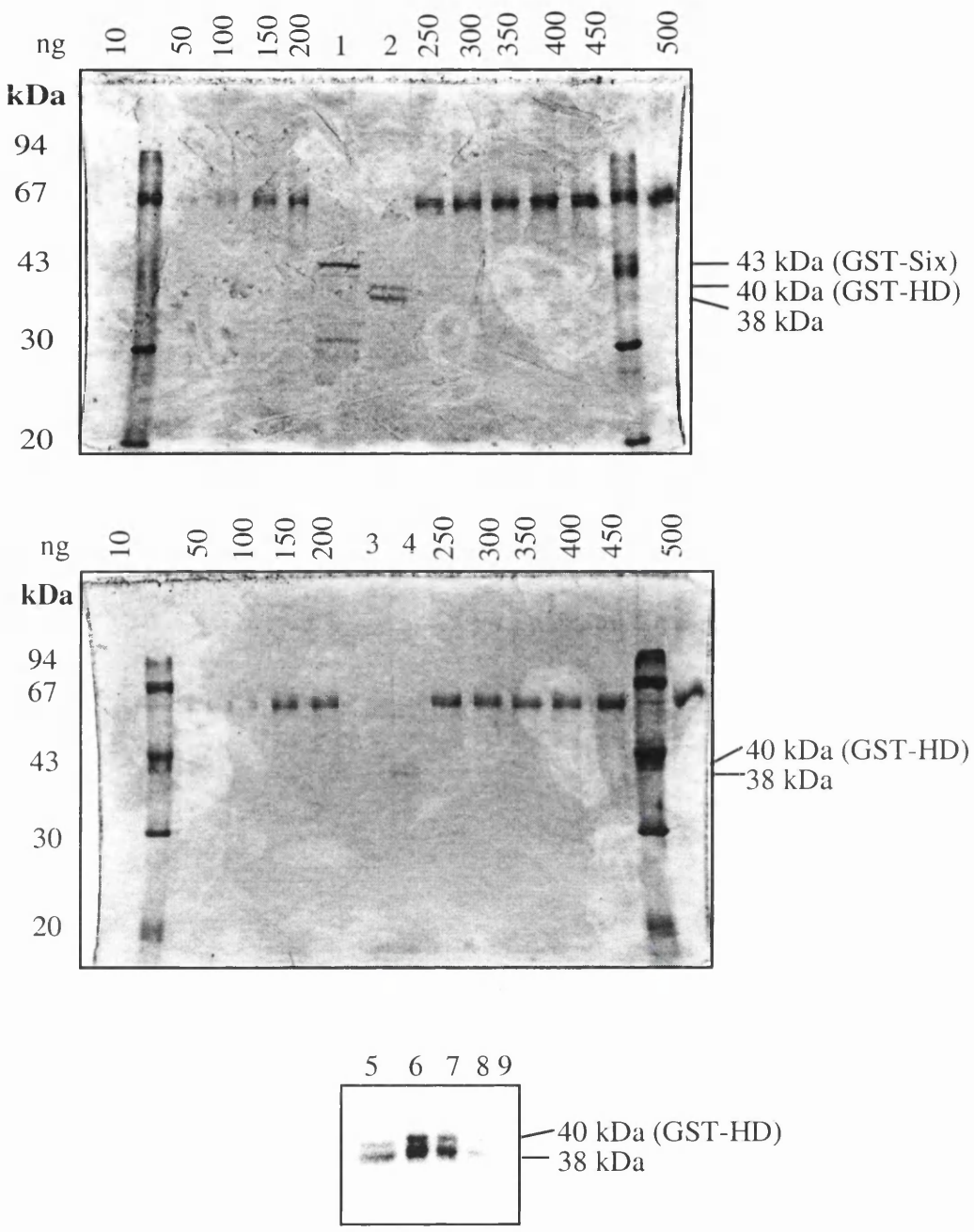

Figure 4.21 SDS-PAGE analysis to determine the concentration of affinity purified recombinant protein

Dilutions of BSA standards in the range of 10-500 ng, were electrophoresed in a 10% (w/v) SDS polyacrylamide gel and stained with Coomassie Blue. These standards were used to determine the concentration of recombinant protein before and after concentration. Lane 1 contains 8 µl (250 ng) affinity purified GST-Six, lanes 2 and 5 contains 8 µl (200 ng) affinity purified GST-HD, lanes 3 and 4 contain 8 µl of GST-Six and GST-HD that have been reduced in volume using Centricon 10 membrane units. Lanes 6, 7, 8 and 9 contain 8 µl acetone precipitated, ethanol precipitated, ethanol precipitated and dialysed and acetone precipitated and dialysed GST-HD, respectively.

that had been precipitated by acetone was about three times more concentrated and the ethanol precipitated sample was about twice as concentrated as the unconcentrated sample. Unfortunately most of the protein was lost during dialysis (figure 4.21). Similar experiments on GST-Six and GST-Six+HD failed to precipitate and thus to concentrate any protein with any degree of efficiency (results not shown).

4.5 Thrombin Cleavage of GST-DMAHP Recombinant Proteins

To obtain recombinant DMAHP protein more representative of native protein, removal of the GST-tag was attempted. A thrombin cleavage site is present in pGEX4T3, upstream of the MCS to allow the GST and the translated product of the subcloned DNA fragment to be separated. Pharmacia Biotech recommend using 10U thrombin per 1 mg of protein to be cleaved. Therefore, 0.01U thrombin were mixed with 1.0 μg affinity purified GST-HD or 1.0 μg affinity purified GST-Six in a total volume of 200 μl and incubated for 24 hours. 15 μl aliquots were taken for SDS-PAGE analysis, every 2 hours. However, due to the low concentration of protein, Coomassie Blue failed to stain the proteins sufficiently to allow visualisation and the proteins were stained using silver nitrate (section 2.3.14.2). This revealed only the presence of the GST-HD and GST-Six proteins plus some faint degradation products that were visible in all samples including those, containing no thrombin (results not shown).

As low concentrations of protein were used, it was thought that small amounts of cleaved protein might not be visible and the experiment was repeated using 0.05U thrombin and 5.0 μg protein. However, even after silver staining the recombinant proteins, no differences could be seen between the samples (results not shown). As the recommended concentration of thrombin had no visible effect on the fusion proteins, the experiment was repeated using 0.1U thrombin (twice that recommended by Pharmacia Biotech) and 5.0 μg protein, but still no cleavage occurred and the experiments were ceased.

4.6 Discussion

The aim of the work described in this chapter was to express and affinity purify recombinant GST-DMAHP proteins that could be used in functional studies of the protein. At the time the work was initiated, nothing was known about the function of DMAHP, although AREC3/Six4 another member of the Six subfamily of proteins had been shown to be a DNA binding protein (Kawakami *et al.*, 1996a).

A bacterial fusion protein expression system was chosen because it allows for fast and efficient purification of protein under nondenaturing conditions. Other DNA binding proteins have been successfully expressed as recombinant fusion proteins and shown to retain their binding activity (Verrijzer and Vandervliet, 1993). Murine members of the *Six* subfamily (*Six2*, *Six3* and *Six5*) have now been expressed as GST fusion proteins and successfully used to investigate the binding activity of these proteins (Kawakami *et al.*, 1996b).

4.6.1 Subcloning of DMAHP Fragments into pGEX4T3

Five mammalian members of the Six subfamily are known and they all share a region of high sequence homology which is composed of the Six box and homeobox. Outside of this region the genes have no sequence homology. Therefore it was believed that the Six box and homeobox encoded the functional domains of the proteins and fragments containing these boxes were subcloned both as a single unit and as two separate units to enable their individual and collective functions to be characterised.

4.6.2 Expression of Recombinant GST-DMAHP Proteins

The growth kinetics of the three fusion proteins (GST-Six, GST-HD and GST-Six+HD) and GST protein were monitored to see the effect of expressing recombinant protein on the *E.coli* cells. Synthesis of recombinant GST protein did not effect the growth rate of the cultures. Both constructs expressing the homeodomain drastically reduced the growth rates of the cultures in which they were being expressed and the construct expressing only the Six domain caused a slight reduction in growth rate. These experiments indicate that the homeodomain is toxic to the cells and that this toxicity causes proteins expressing the homeodomain to be degraded. A GST-DMAHP protein produced by the subcloning of the RT-PCR product present in pMOS-DMAHP-ABC (figure 4.1) into pGEX-2T (GST-DMAHP-ABC), caused reduced growth of *E.coli* cultures expressing it (Winchester, 1997). However, in contrast to the homeodomain containing proteins expressed for this thesis, GST-DMAHP-ABC was not degraded, but was expressed very slowly. Optimum expression was obtained after 21 hours of induction by 1 mM IPTG at 37°C.

GST-HD was never expressed in high enough concentrations to be visible on a Coomassie Blue stained gel, in the presence of bacterial proteins, as it was masked by a bacterial protein of the same weight. However, a 30 kDa GST fusion protein could be seen after induction at 30 °C. After affinity purification the bulk of the recombinant

GST-HD protein was detected as a doublet, with the two bands of 38 and 40 kDa staining with equal intensity. Both bands were detected by GST antiserum and R254 DMAHP antiserum indicating that they are GST fusion proteins containing the homeodomain of DMAHP. The 30 kDa protein was detected by GST antiserum, but not by R254 DMAHP antiserum, indicating that it consisted mainly of recombinant GST. It was assumed that the 40 kDa protein, represented the full length expressed protein. The 38 kDa protein may be a truncated version of this protein or alternative folding may have caused half of the protein sample to electrophorese slightly faster than the other half during SDS-PAGE.

Expression of GST-Six did not significantly reduce the growth of the culture in which it was being expressed and relatively large amounts of this protein were produced. However, even when the recombinant GST-Six was induced at 30°C it was insoluble. This is a common problem with recombinant proteins produced in bacteria and may be due to the high concentration of the protein within the bacterial cell or to the particular amino acid sequence of the protein causing it to aggregate into inclusion bodies (Schein, 1989). Inclusion body solubilisation of GST-Six appeared to have been successful as the 42 kDa protein could be visualised in the solubilised pellet fraction. However, after affinity purification only a small fraction of the protein was present in the eluant suggesting that the Six domain part of the fusion protein had interacted with the matrix and could therefore not be eluted using reduced glutathione. Alternatively, the majority of the protein had precipitated in the 1 x PBS, after dialysis and was simply being pelleted during the centrifugation stages of affinity purification. GST-Six was detected by GST antiserum and was not detected by R254 DMAHP antiserum, indicating that it was indeed a GST fusion protein. No Six domain antiserum was available to confirm that the Six domain was present in this recombinant protein, but sequence analysis of the pGEX.Six construct indicated that this should be so.

Like GST-HD, GST-Six+HD was never expressed in high enough concentrations to be visible by Coomassie Blue staining, in the presence of bacterial proteins, as it was masked by a bacterial protein of the same weight. GST-Six+HD was detected in the pellet fraction of the crude cell lysate, produced during a time course experiment, by R254 DMAHP antiserum. The amount of recombinant protein detected decreased with time, indicating that the protein was being degraded. When GST-Six+HD was induced in a 500 ml culture at 30°C for 1 hour the protein was detected in the supernatant fraction of the crude cell lysate. This may have been due to either the reduced temperature increasing the solubility of the protein or to sonication being a more efficient way of lysing bacterial cell walls than repeated freeze/thawing.

GST-Six+HD was detected as a doublet in an SDS polyacrylamide gel. The two bands of 50 and 52 kDa were detected by R254 DMAHP antiserum, indicating that they contained the homeodomain of DMAHP. They were visible with equal intensity when stained with Coomassie Blue. It was assumed that the larger protein (52 kDa), represented the full length expressed protein and that the 50 kDa protein may be either a shorter version of this protein or alternative folding may have caused half of the protein sample to electrophorese slightly faster than the other half during SDS-PAGE. It was interesting that the two proteins expressing the homeodomain were detected as doublets, in SDS polyacrylamide gels, with the upper and lower bands being visualised with equal intensity when stained with Coomassie Blue. In both cases the upper band was two kDa larger than the lower band. It is possible, that a site susceptible to proteolytic cleavage exists at the C-terminus of the homeodomain or that the structure of the homeodomain, when expressed as a recombinant protein in bacteria, allows it to fold in two different ways. It is also possible that the lower band in each doublet represented truncated expressed protein. A GST-DMAHP protein produced by the subcloning of the RT-PCR product present in pMOS-DMAHP-ABC into pGEX-2T was always detected as a single band. However, the predicted molecular weight of this protein was 71 kDa, but was detected as 37 kDa (Winchester, 1997). It was thought to be truncated at the C-terminus which would mean that translation was ceasing at about the same position as in the homeodomain proteins that were expressed for this thesis, if the smaller bands did indeed represent truncated expressed protein. GST fusion proteins made containing murine Six2, Six3, Six4/AREC3 and Six5/Dmahp homeodomains were all detected as single bands on SDS polyacrylamide gels (Kiyoshi Kawakami, personal communication).

4.6.3 Concentrating Recombinant GST-DMAHP Proteins

Although the affinity purified GST-DMAHP proteins produced were relatively pure, it was only possible to obtain them at between 15 and 30 ng/μl of eluted protein. This is sufficiently concentrated for gel retardation assays, but leaves little scope for carrying out titration experiments to determine the optimum ratio of DNA:protein. This information would be beneficial when calculating the amount of protein to use in the whole genome PCR screen (chapter 6). Therefore, ultrafiltration was used in the form of Centricon 10 columns, to concentrate GST-Six and GST-HD. Unfortunately, the proteins appeared to bind permanently to the membranes, which is a common problem when using proteins with binding activity (Martin Boocock, personal communication). Although this was disappointing with regard to concentrating the proteins, it was encouraging from the point of view that proteins showed binding

activity. As GST-Six and GST-HD both attached permanently to the membrane, when GST-Six+HD was purified it was not concentrated by ultrafiltration.

As the recombinant proteins could not be purified by ultrafiltration, precipitation by organic solvents was tried. Both acetone and ethanol were used although acetone is thought to carry less risk of denaturing the protein. Only GST-HD was successfully concentrated in this manner and about 70% of the protein was lost when using acetone, which was more efficient than ethanol. Ammonium sulphate was also considered a possibility for precipitating the proteins. However, for this to work with any degree of efficiency an initial concentration of 1 μg/μl is required and so this was not attempted. These experiments indicate that the recombinant proteins are at least partially soluble in 50% (v/v) acetone and ethanol. Lyophilisation was also considered as a possibility for concentrating the recombinant proteins. However previous experiences in our laboratory, of lyophilising GST fusion proteins, had shown that only 25% of the protein which had been lyophilised was recovered. Therefore, it was decided to proceed with gel retardation assays using unconcentrated recombinant protein and to test the concentrated GST-HD to see if it greatly improved the binding activity of this particular recombinant protein (chapter 5).

4.6.4 Thrombin Cleavage of Recombinant GST-DMAHP Proteins

Although DNA-protein binding experiments have been successfully carried out using complete GST fusion proteins (Verrijzer *et al.*, 1992), it was decided to attempt to remove the GST moiety from GST-HD and GST-Six to obtain more authentic proteins. However, despite using twice the recommended amount of thrombin, no cleavage occurred. It is possible that the GST-HD and GST-Six fusion proteins folded in such a way as to make the cleavage site inaccessible to the thrombin. A GST-DMAHP protein produced by the subcloning of the RT-PCR product present in pMOS-DMAHP-ABC into pGEX-2T also failed to cleave using thrombin (Winchester, 1997) and the inefficient cleavage of several other GST fusion proteins has been reported (Guan and Dixon, 1991). Guan and Dixon (1991) improved the cleavage efficiency of several fusion proteins by introducing a glycine-rich linker, immediately following the thrombin cleavage site in the pGEX vector. It is thought, that the tandem glycine residues alter the structure at the protease cleavage site within the fusion protein, thus making the site more accessible to thrombin. To address the common problem of inefficient cleavage by thrombin, Pharmacia Biotech have now introduced a new pGEX vector that contains a specific cleavage site for a genetically engineered protease, PreScission™ Protease, which they report greatly enhances cleavage efficiency.

As thrombin cleavage of GST-Six and GST-HD failed, no attempt was made to cleave GST-Six+HD once it was affinity purified. Unfortunately no substrate was available on which to test the activity of the thrombin. All further experiments were carried out using complete GST fusion proteins.

4.7 Conclusion

Three recombinant GST-DMAHP fusion proteins, containing the Six domain, the homeodomain and both the Six domain and homeodomain together, were successfully overexpressed in a bacterial system. The recombinant fusion proteins were affinity purified with glutathione as the affinity ligand.

Chapter

5

Characterisation of the DNA Binding Properties of Recombinant DMAHP Proteins

5.1 Introduction

The recombinant GST fusion proteins (GST-Six, GST-HD and GST-Six+HD) described in the previous chapter were used in gel retardation assays to investigate putative DNA binding targets of DMAHP and the functions of its two conserved domains. Double stranded oligonucleotides representing putative DNA binding sites were utilised as DNA fragments in the assays.

AREC3/Six4 binds to the Na^+ K^+ ATPase α 1 subunit gene regulatory element (ARE) and is a cell-type specific factor involved in the regulation of this gene (Kawakami *et al.*, 1996a). The AREC3/Six4 binding site within the ARE (GGTGTCAGGTTGC) differs from all previously reported homeodomain binding sites, the majority of which include a core tetranucleotide ATTA (Gehring *et al.*, 1994). The amino acid at position 50 of the classical homeodomain specifies the two bases immediately 5' to the core sequence. DMAHP has a lysine at position 50 which would indicate that its binding site might be GGATTA. However, it is known that arginine at position 5 of the homeodomain which is conserved in 97% of known homeodomains binds to the core sequence (ATTA) (Gehring *et al.*, 1994). DMAHP and AREC3/Six4 have a valine at this position and other members of the Six subfamily have serine or threonine (section 1.8.2). Therefore, it is likely that the Six subfamily of homeodomain proteins bind to their target DNA at an entirely different recognition site to other homeodomain proteins. The Six domain may play an important role in modulating binding specificity, either as a DNA binding domain or a protein binding domain or both.

It was hypothesised that *DMAHP* regulates *DMPK* (the protein kinase gene present at the DM locus) as prior to the start of this project both genes had been shown to be expressed in a similar range of tissues and levels of expression of *DMPK* had been shown to be altered in DM patients (Boucher *et al.*, 1995; Carango *et al.*, 1993; Fu *et al.*, 1993; Krahe *et al.*, 1995a; Sabourin *et al.*, 1993; Wang *et al.*, 1995). A site similar to the AREC3/Six4 binding site was identified in the promoter region of human *DMPK*, along with a GGATTA homeodomain consensus site which is conserved in both mouse and human.

5.2 Subcloning of DNA Fragments into pBluescript® SK(+)

Three putative DNA binding sites of DMAHP were identified: 1) The 13 bp AREC3/Six4 binding site, GGTGTCAGGTTGC (Suzuki-Yagawa *et al.*, 1992). 2) A Bestfit analysis of the AREC3/Six4 binding site to the promoter region of *DMPK* identified CCCCTCAGGTTCC (nucleotides 733-745, accession number L08835) as

the most similar in this region. This sequence was termed ARE-like. 3) A GGATTA consensus site in the promoter region of *DMPK* (nucleotides 1323-1328, accession number L08835). 40 bp double stranded oligonucleotides representing the genomic sequences containing the putative sites and flanking regions were produced. A mutant ARE sequence, GGTGT**G**AGGTTGC, that differed by a single base pair (highlighted in bold) from the true ARE sequence and had been shown not to bind AREC3/Six4 (Kawakami *et al.*, 1996a) was also made as a negative control. The oligonucleotides (sequences are shown in table 2.4) were subcloned into pBluescript® SK(+) (section 2.3.17.1).

5.3 Gel Retardation Assays Using Affinity Purified Recombinant Protein

Gel retardation assays were performed using affinity purified GST, GST-Six, GST-HD and GST-Six+HD and radiolabelled DNA fragments designated, ARE, AREmut, ARE-like and GGATTA (section 2.3.17). Plasmids containing the putative binding sites were digested with *Hind* III and *Xba* I to release a 66 bp fragment, containing the 40 bp genomic sequence and 26 bp of plasmid sequence, from pBluescript SK(+). Both pBluescript SK(+) and the DNA fragment containing the putative binding site were radiolabelled (section 2.3.17.2). The pBluescript SK(+) DNA acted as a non-specific DNA competitor. Reactions contained 6 pmoles (10 µl of eluant from the affinity purification process, section 4.3.2) of recombinant protein and 3 fmoles of DNA fragment and were carried out in both the presence and absence of 0.01 µg/µl of poly[dI/dC].poly[dI/dC], which also acted as a non-specific competitor. It was decided to add 10 µl of undiluted affinity purified recombinant protein to a reaction mix with a final volume of 15 µl because the concentration of the affinity purified GST-DMAHP recombinant proteins was relatively low (15-30 ng/µl ≅ 0.6 µM). This made the final concentration of the binding buffer 33 mM Tris-HCl (pH 8.0), 80 mM NaCl, 7 mM DTT, 0.7 mM PMSF which is similar to buffers used for other homeodomain protein gel retardation assays (Affolter *et al.*, 1990; Wilson *et al.*, 1993). Affinity purified GST recombinant protein was diluted in 1.5 x binding buffer so that it too was at a concentration of 0.6 µM. Reactions were electrophoresed in a standard 8% (w/v) low ionic strength or 1 x Tris/glycine, non-denaturing polyacrylamide gel (section 2.3.17.4).

5.3.1 Binding Reactions Analysed in a Standard Low Ionic Strength Non-Denaturing Polyacrylamide Gel

GST-Six+HD was combined with each of the radiolabelled DNA fragments and the reactions were analysed in a standard low ionic strength non-denaturing polyacrylamide gel. Binding of the recombinant protein to the DNA fragments was not detected under these conditions (figure 5.1a). Therefore, a 1 x Tris/glycine gel and buffer system was utilised as this system had been reported to stabilise certain protein/DNA complexes that had proven to be unstable under standard conditions (Bednarz *et al.*, 1990).

5.3.2 Binding Reactions Analysed in 1 x Tris/Glycine Non-Denaturing Polyacrylamide Gels

GST-Six+HD was combined with each of the radiolabelled DNA fragments and the reactions were analysed in a 1 x Tris/glycine non-denaturing polyacrylamide gel. Two retarded complexes were detected in the lanes containing the ARE fragment (C1 and C2), but not in any of the other lanes (figure 5.1b). The binding activities of all 4 recombinant proteins (GST, GST-Six, GST-HD and GST-Six+HD) were tested with the ARE fragment. GST-Six-HD formed 2 complexes and GST-HD formed a single complex (C3). Neither GST nor GST-Six showed any binding activity with the ARE. A lower yield of complex was formed by GST-HD than GST-Six+HD and the complex was of an intermediate mobility (figure 5.1c). The binding activity of GST-HD was tested using ARE, AREmut, ARE-like and GGATTA. Binding was only detected with ARE (C3, figure 5.1d). The presence of 0.01 μg/μl of poly[dI/dC].poly[dI/dC] as a non-specific DNA competitor had no effect on the formation of the retarded complexes. GST-Six and GST showed no binding activity when tested with AREmut, ARE-like or GGATTA (results not shown).

The experiments described above indicate that both GST-Six+HD and GST-HD display specific DNA binding activity. However, the extent of binding to the ARE fragment was not substantial. Less than 10% (w/w) of the labelled fragment was bound by the recombinant proteins. This may have been due to the relatively low concentration of protein used. Attempts to concentrate the recombinant proteins met with little success (section 4.4.2). However, GST-HD was successfully precipitated by ethanol and acetone and then resuspended in 1.5 x binding buffer to give concentrations of 1.2 μM (50 ng/μl) and 1.8 μM (75 ng/μl) respectively (section 4.4.2b). The binding activity of concentrated GST-HD was tested with the ARE fragment; it was shown to have lost all binding activity (figure 5.2).

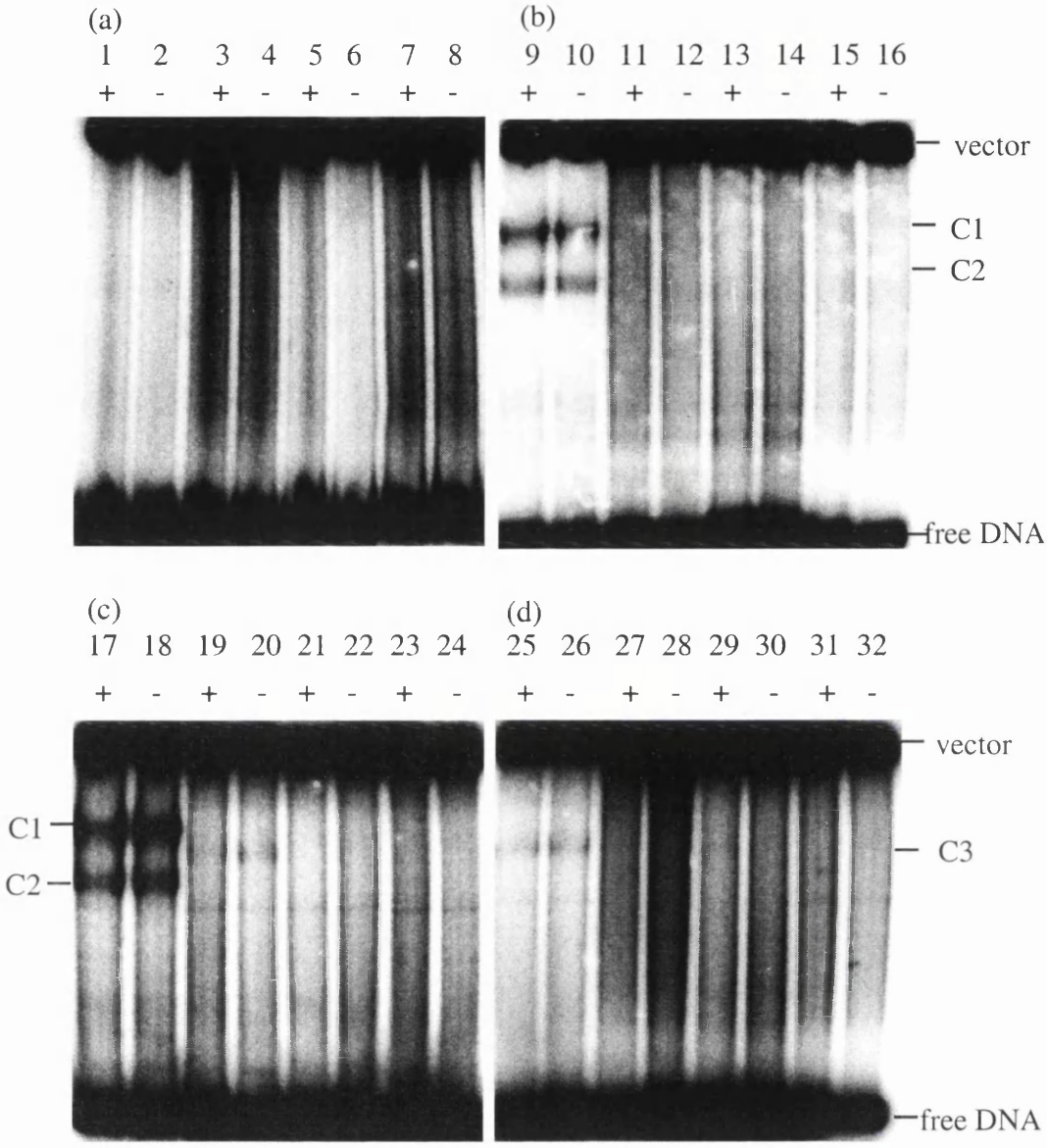

Figure 5.1 Gel retardation assays using affinity purified GST-DMAHP recombinant proteins

Reactions contain 6 pmoles of affinity purified recombinant protein and 3 fmoles of labelled DNA fragment. All reactions were performed in the presence (+) and absence (-) of 0.01μg/μl poly[dI/dC].poly[dI/dC]. GST-Six+HD was tested with: ARE (lanes 1, 2, 9 and 10) ARE-like (lanes 3, 4, 11 and 12), GGATTA (lanes 5, 6, 13 and 14), ARE-mut (lanes 7, 8, 15 and 16). The reactions were anlaysed in a standard low ionic strength gel and buffer system (a) or a Tris/glycine gel and buffer system (b). GST-Six+HD (lanes 17 and 18), GST-HD(lanes 19 and 20), GST-Six (lanes 21 and 22), GST (lanes 23 and 24) were tested with ARE in a Tris/glycine system (c). GST-HD was tested with ARE (lanes 25 and 26), ARE-like (lanes 27 and 28), GGATTA (lanes 29 and 30), AREmut (lanes 31 and 32) in a Tris/glycine system (d).

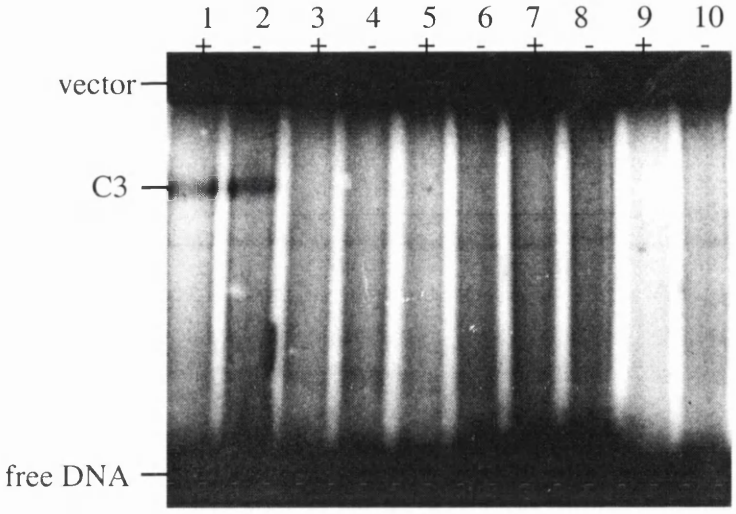

Figure 5.2 Comparison of binding activity of precipitated and non-precipitated GST-HD

10 µl of affinity purified, unprecipitated (lanes 1and 2), acetone precipitated (lanes 3 and 4), ethanol precipitated (lanes 5 and 6), ethanol precipitated and dialysed (lanes 7 and 8) and acetone precipitated and dialysed (lanes 9 and 10) GST-HD were tested against ARE in the presence (+) and absence (-) of 0.01 µg/µl poly[dI/dC].poly[dI/dC].

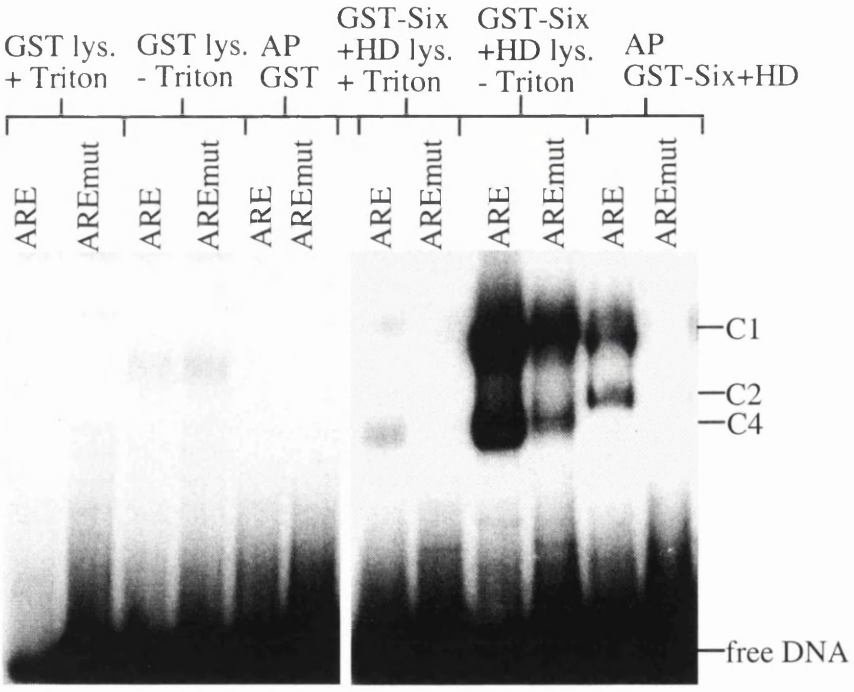

Figure 5.3 Gel retardation analysis using purified ARE fragment

The binding activities of 10 µl of 1/10 dilutions of cell lysate (lys.) prepared with (+ Triton) and without (- Triton)1% (v/v) Triton X-100, from bacteria expressing GST or GST-Six+HD and affinity purified (AP) GST and GST-Six+HD, were tested with 3 fmoles of gel purified ARE and AREmut.

5.3.3 Gel Retardation Assay Using Purified ARE Fragment

The homeodomain of *Drosophila* sine oculis was reported to bind the core sequence (C/T)GATAC present within the 3.1 kb pUC119 plasmid (Hazbun *et al.*, 1997). This sequence is also represented 4 times in the related plasmid pBluescript SK(+), which was present as a non-specific competitor in the gel retardation experiments described above. Therefore, to determine the effect of pBluescript SK(+) in gel retardation assays the putative binding sites were separated from the vector sequence. The 66 bp fragment released from pBluescript SK(+) by digestion with *Hind* III and *Xba* I was purified from a 3.5% (w/v) NuSieve agarose gel using a Qiagen gel extraction DNA purification kit. The purified fragment was radiolabelled (section 2.3.17.2, with the exception that purified fragment plus 1 x Life Technologies Ltd React 2 was used in place of the restriction digest reaction). The binding activities of affinity purified GST-Six+HD and GST were tested with gel purified ARE and AREmut. 0.01 μg/μl of poly[dI/dC].poly[dI/dC] was used as a non-specific DNA competitor, as it was in all further gel retardation assays. No increase in the amount of either DNA fragment binding to affinity purified GST-Six+HD was detected and affinity purified recombinant GST bound to neither ARE nor AREmut (figure 5.3). Therefore, it was decided that it was not necessary to separate the putative binding sites from the vector sequence prior to a gel retardation assay.

5.4 Gel Retardation Assays Using Crude Cell Lysate Containing Recombinant Protein

As the binding activities of affinity purified DMAHP fusion proteins were relatively low, the binding activities of GST, GST-Six+HD and GST-HD in crude cell lysate were investigated. GST-Six was unsuitable for this experiment due to its insolubility in the cell lysate (section 4.3.2c). 500 ml cultures containing pGEX4T3, pGEX.Six+HD or pGEX.HD were grown and recombinant protein expression induced by the addition of 1 mM IPTG (section 2.3.12.3). The cells were lysed by sonication at room temperature and either 1% (v/v) Triton X-100 was added and incubated at room temperature for 30 minutes before the sample was centrifuged and the supernatant snap frozen, or the sample was centrifuged and the supernatant snap frozen immediately. Triton X-100 was added to samples prior to affinity purification to aid the solubilisation of the fusion proteins. However, there was concern that the presence of detergent in the crude cell lysate would disrupt DNA-protein complexes formed by protein present in the cell lysate.

Crude cell lysate from bacteria expressing GST-Six+HD was diluted 1/10 with 1.5 x binding buffer. The binding activity of 10 µl of diluted cell lysate containing overexpressed GST-Six+HD, both with and without 1% (v/v) Triton X-100, was tested with gel purified ARE and AREmut. The crude cell lysate in the absence of Triton X-100 bound strongly to the ARE and also bound with a lower affinity to AREmut. The crude cell lysate in the presence of Triton X-100 bound very weakly to the ARE and no binding was observed to AREmut (figure 5.3). It was observed that the lower complex, formed by the crude cell lysate (C4) had an apparent molecular weight that was less than that of the lower complex formed by affinity purified GST-Six+HD (C2). To demonstrate that it was the overexpressed GST-Six+HD forming the retarded complexes with the DNA and not another protein present in the cell lysate, the binding activity of cell lysate from bacteria expressing recombinant GST was investigated. GST cell lysate in the presence of Triton X-100 did not bind to either of the fragments. In the absence of Triton X-100 it bound weakly to both ARE and AREmut, (figure 5.3). This binding was very weak and the complex was of a different molecular weight to the complexes detected when using unpurified GST-Six+HD and so it was concluded that the complexes detected in the presence of crude cell lysate from bacteria expressing GST-Six+HD were indeed formed by the GST-Six+HD.

The binding activities of both GST-Six+HD and GST-HD, from crude cell lysates prepared without the addition of Triton X-100, were investigated using the 4 putative DNA binding sites. The reactions were carried out in the presence of radiolabelled DNA fragment and vector (pBluescript SK(+)), plus 0.01 µg/µl of poly[dI/dC].poly[dI/dC]. Interestingly, in the presence of pBluescript SK(+), GST-Six+HD bound strongly to the ARE, forming 2 retarded complexes, but bound to AREmut with much less affinity than in the absence of pBluescript SK(+). The upper complex formed with the ARE contained about 3 times as much DNA as the lower complex. No complexes were detected with ARE-like or GGATTA (figure 5.4). The binding activity of GST-HD appeared unaltered when present in crude cell lysate. Weak binding was observed to the ARE, but not to AREmut, ARE-like or GGATTA (figure 5.4).

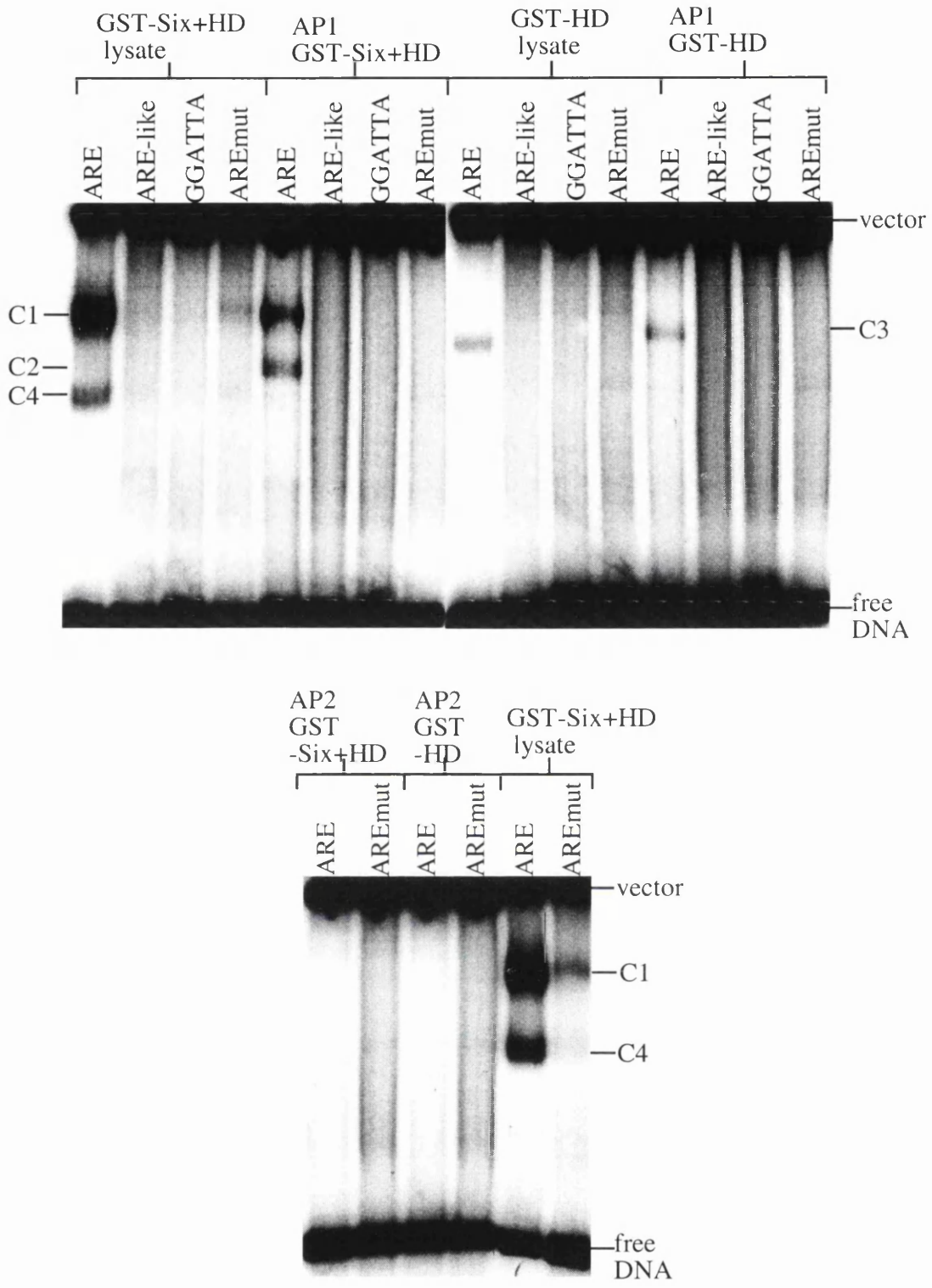

Figure 5.4 Comparison of the binding activities of affinity purified and unpurified GST-DMAHP recombinant proteins

The binding activities 10 µl of 1/10 dilution of cell lysates containing GST-Six+HD or GST-HD were compared to the binding activities of 6 pmoles of the same recombinant proteins that had been affinity purified either with (AP1) or without (AP2) the addition of 1% (v/v) Triton X-100. 3 fmoles of ARE, ARE-like, GGATTA and AREmut were used as probes.

5.5 Investigations into the Cause of the Increased Binding Activity of Unpurified GST-Six+HD

Several explanations for the increase in binding activity, particularly the increase in DNA present in the upper complex, when GST-Six+HD remained in the crude cell lysate were investigated. These are as follows:

1) Triton X-100 is present in affinity purified GST-Six+HD at a low concentration causing the DNA-protein complex to partially dissociate.

2) Affinity purification is inefficient, causing a reduction in the concentration of GST-Six+HD in the purified eluant compared to the unpurified cell lysate.

3) The protein is highly susceptible to degradation in its affinity purified form.

4) Affinity purification modifies the GST-Six+HD in such a way as to reduce its binding activity.

5) A cofactor, similar to a natural cofactor present in the human cell, is present in the *E.coli* cell lysate that enhances binding activity.

5.5.1 Affinity Purification of Recombinant Proteins in the Absence of Triton X-100

The presence of Triton X-100 in the crude cell lysate greatly reduced the formation of DNA-protein complexes. Therefore, it was possible that trace amounts of Triton X-100 in affinity purified GST-Six+HD partially disrupted the formation of the DNA-protein complexes. Therefore, GST-Six and GST-Six+HD were affinity purified (section 2.3.12.5) without the post-sonication addition of 1% (v/v) Triton X-100. The yield of affinity purified protein was reduced from between 15 and 30 ng/µl to 10 ng/µl. This was presumably due to the reduced solubility of the recombinant proteins in the absence of Triton X-100. GST-Six+HD and GST-HD purified in this manner, showed no binding activity when tested with ARE and AREmut (figure 5.4). This was likely to be due to the decrease in the concentration of the recombinant proteins. This experiment indicated that the increase in binding activity was not due, solely, to the absence of Triton X-100 in the binding reaction.

5.5.2 Analysis to Determine Relative Amounts of Binding Activity in Affinity Purified Recombinant Proteins and Crude Cell Lysates

As the lower binding activity of affinity purified GST-Six+HD did not appear to be due to the presence of Triton X-100 in the binding reaction other factors were investigated. It was hypothesised that either total recombinant protein or the binding activity of the recombinant protein was reduced by affinity purification. 10 µl of affinity purified eluant represents 0.17% (v/v) of total eluant, whereas 10 µl of a 1/10 dilution of crude cell lysate represents only 0.005% (v/v) of total cell lysate. Therefore if the affinity purification were 100% efficient, reactions containing affinity purified protein would contain 34 times as much recombinant protein as those performed with crude cell lysate.

A titration assay was performed to determine the concentration of crude cell lysate from bacteria expressing GST-Six+HD required to produce the level of binding observed when using 10 µl (6 pmoles) of affinity purified GST-Six+HD. Cell lysate from bacteria expressing GST-Six+HD was diluted 1/10, 1/20, 1/40, 1/80, 1/160 and 1/320 in 1.5 x binding buffer. 10 µl of each dilution were tested in a binding assay using ARE. The intensity of the 2 complexes gradually decreased, until at the dilution of 1/160 the level of binding was approximately that seen with affinity purified GST-Six+HD (figure 5.5). A 1/10 dilution of GST-HD cell lysate gave a level of binding equivalent to 10 µl (6 pmoles) of affinity purified GST-HD (figure 5.4).

5.5.3 Western Blot Analysis of Affinity Purified and Unpurified GST-Six+HD and GST-Six+HD

To determine the efficiency of the affinity purification process it is necessary to know the concentrations of the recombinant fusion proteins in both their affinity purified forms and in crude bacterial cell lysate. However, it was not possible to determine the concentration of the affinity purified recombinant fusion proteins using standard assays as they were not present at a high enough concentration. The concentrations of the unpurified recombinant fusion proteins could not be determined by such an assay due to the presence of bacterial proteins. Therefore, the relative concentrations of affinity purified and unpurified GST-Six+HD and GST-HD were determined by western blot analysis. Serial dilutions of cell lysates from bacteria expressing GST-Six+HD and GST-HD and undiluted affinity purified recombinant proteins were electrophoresed in 10% (w/v) SDS polyacrylamide gels. Protein was transferred to Hybond™-N ECL nitrocellulose membrane and probed with R254 DMAHP antiserum (section 2.3.16). A 20 minute exposure to detection solution was required

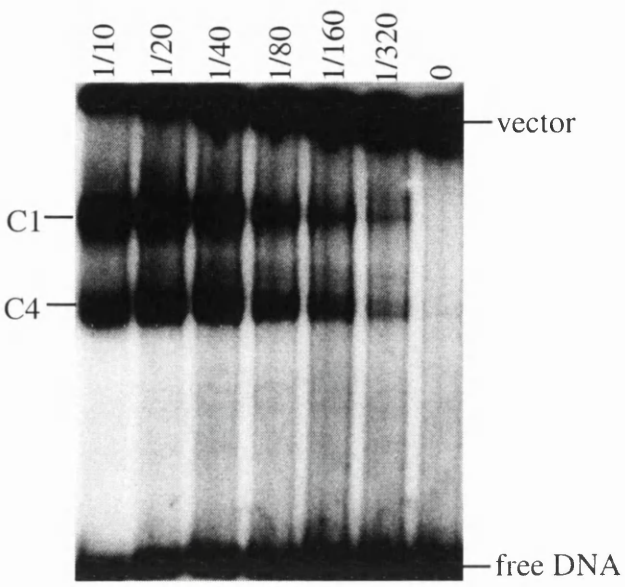

Figure 5.5 Gel retardation analysis of a serial dilution of GST-Six+HD cell lysate

Serial dilutions of cell lysate expressing GST-Six+HD were made. The binding activity of 10 µl of each dilution was tested with 3 fmoles ofARE. As the concentration of cell lysate decreased the intensities of both C1 and C4 decreased.

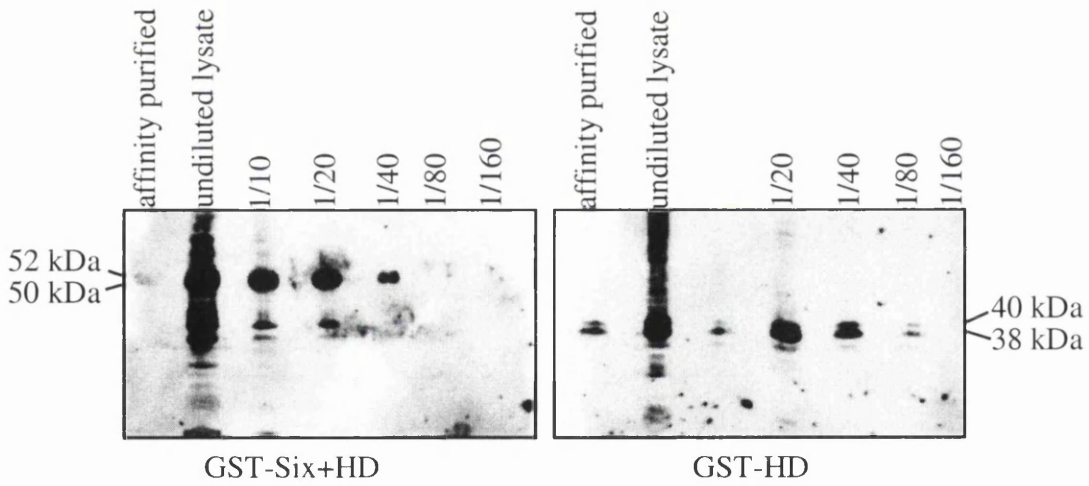

Figure 5.6 Western blot analysis of serial dilutions of GST-Six+HD and GST-HD cell lysates

10 µl of affinity purified GST-Six+HD and GST-HD, undiluted cell lysate containing each of the recombinant proteins and dilutions of each of the recombinant protein cell lysates were electrophoresed in 10% (w/v) SDS polyacrylamide gels. The protein was transferred to Hybond-N ECL nitrocellulose membrane and probed with R254 DMAHP antiserum.

to visualise 6 pmoles of affinity purified GST-Six+HD (membranes are normally exposed for a maximum of 5 minutes). The amount of GST-Six+HD and GST-HD, detected in the cell lysate gradually decreased as the cell lysate became less concentrated. After a 20 minute exposure, GST-Six+HD was no longer visible at a dilution of 1/80 (figure 5.6). Therefore, R254 DMAHP antiserum has the equivalent affinity for approximately 10 µl of a 1/60 dilution of crude cell lysate from bacteria expressing GST-Six+HD as for 10 µl (6 pmoles) of undiluted affinity purified GST-Six+HD. GST-HD could still be detected in a 1/80 dilution of crude cell lysate from bacteria expressing this recombinant protein, but could no longer be detected at a concentration of 1/160. It was estimated that R254 DMAHP antiserum would have an equivalent affinity for GST-HD in a 1/60 dilution of crude cell lysate as 10 µl (6 pmoles) of undiluted affinity purified GST-HD. Although cell lysates from bacteria expressing GST-Six+HD needed to be diluted to the same degree as those from bacteria expressing GST-HD, for R254 DMAHP antiserum to bind with the same affinity to both unpurified and 10 µl (6 pmoles) of affinity purified recombinant protein, it had a stronger overall affinity for GST-HD.

If the antibodies bound to affinity purified and unpurified recombinant proteins to the same extent then the western blot analysis suggests that affinity purification, purified only 0.5% of the overexpressed GST-Six+HD. This would mean that 10 µl of crude cell lysate from bacteria expressing the recombinant proteins would contain approximately 20 µg of GST-Six+HD and 15 µg of GST-HD (approximately 350 pmoles) and this mass of protein should have been detected by SDS-PAGE analysis. No overexpressed recombinant proteins were detected by SDS-PAGE analysis of the crude cell lysates (section 4.3.1).

Recombinant GST-Six+HD in a 1/160 dilution of crude cell lysate from bacteria expressing it, bound to the ARE with the same affinity as undiluted purified GST-Six+HD. However, undiluted purified recombinant protein was detected by R254 DMAHP antiserum, whereas the 1/160 diluted unpurified GST-Six+HD was not. GST-HD in a 1/10 dilution of crude cell lysate from bacteria expressing it, bound the DNA with the same affinity as affinity purified GST-HD. However, R254 DMAHP antiserum had a stronger affinity for the recombinant protein in the 1/10 dilution of cell lysate than for the undiluted purified protein. These results indicate that the increase in the binding activity of GST-Six+HD when unpurified may not have been simply due to an increase in concentration of recombinant protein. They also infer the possibility that affinity purification of both GST-Six+HD and GST-HD altered their conformation in such a way as to reduce antigenicity.

5.5.4 Binding Activity of the Cell Lysate

The possibility that a cofactor, present in the crude cell lysate was responsible for the enhanced binding activity of GST-Six+HD was investigated. A gel retardation assay was performed with 9 µl of affinity purified GST-Six+HD (5.4 pmoles) and 1 µl of undiluted crude GST cell lysate with ARE as the DNA bait. Control experiments included reactions with 10 µl of affinity purified GST-Six+HD (6 pmoles), 10µl of 1/10 dilution of crude cell lysate from bacteria expressing recombinant GST-Six+HD and crude cell lysate from bacteria expressing recombinant GST. All reactions were repeated using AREmut as the DNA fragment. The reaction containing 5.4 pmoles of affinity purified protein and 1µl of recombinant GST containing cell lysate resulted in about 80% (w/w) of the ARE DNA forming a single retarded complex (C5). The complex had a mobility consistent with a higher molecular weight than the upper complex formed by unpurified and affinity purified GST-Six+HD. Unpurified GST, in the absence of affinity purified GST-Six+HD produced no retarded complexes (figure 5.7). Due to the toxicity of GST-Six+HD to bacterial cell cultures (section 4.3.1c), the concentration of all proteins (with the exception of GST-Six+HD) was higher in the GST expressing cell lysate than in the GST-Six+HD expressing cell lysate. After induction with 1 mM IPTG, the cell density of the GST expressing culture was 1.3 times higher than that expressing GST-Six+HD (section 2.3.12.3). This experiment strongly suggests that a cofactor was present in the crude cell lysate that enhanced the binding of GST-Six+HD to the ARE.

5.6 Specific Binding of GST-HD to the ARE

Affinity purified and unpurified GST-HD bound to the ARE with low affinity. To determine the specificity of this binding, gel retardation assays were carried out using affinity purified GST-HD and ARE in the presence of a 10 fold molar excess of unlabelled competitor DNA. The competitor DNA fragments used were ARE and AREmut. An excess of unlabelled ARE successfully competed for the recombinant protein. However, an excess of AREmut had no effect on the binding of GST-HD to the labelled ARE (figure 5.8) indicating that although only a small proportion of ARE (less than 1% (w/w)) bound to the recombinant protein, this binding was specific.

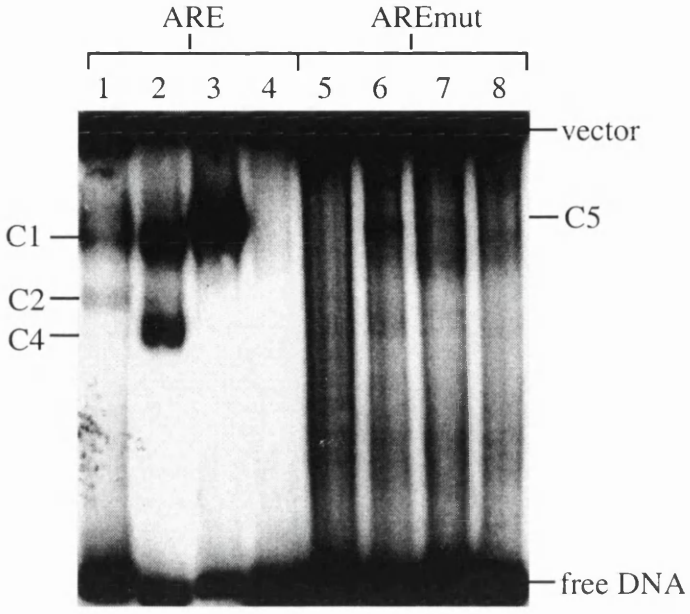

Figure 5.7 Gel retardation analysis of affinity purified GST-Six+HD in the presence and absence of cell lysate

The binding activity of affinity purified GST-Six+HD combined with cell lysate from bacteria expressing recombinant GST was investigated by gel retardation analysis with 3 fmoles of ARE and AREmut. Lanes 1 and 5 contain 10 µl (6 pmoles) of affinity purified GST-Six+HD, lanes 2 and 6 contain 10 µl of a 1/10 dilution of GST-Six+HD cell lysate, lanes 3 and 7 contain 9 µl (5.4 pmoles) of affinity purified GST-Six+HD and 1 µl of undiluted GST cell lysate and lanes 4 and 8 contain 10 µl of a 1/10 dilution of GST cell lysate.

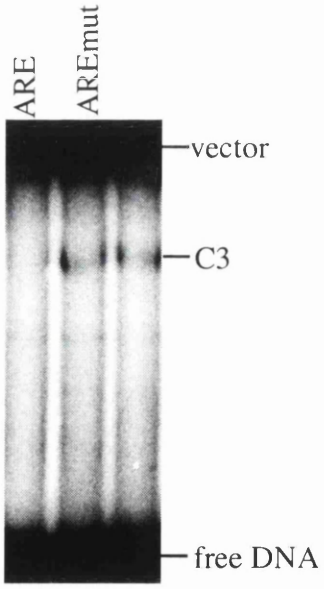

Figure 5.8 Competition asasy of GST-HD

A gel retardation asasy was performed using 6 pmoles of affinity purified GST-HD and 3 fmoles of radiolabelled ARE. Lane 1 contains 30 fmoles of unlabelled ARE, lane 2, 30 fmoles of unlabelled AREmut and lane 3, no competitor.

5.7 Discussion

When this work was instigated no binding studies involving DMAHP/Six5 had been reported. Of the other mammalian members of the Six subfamily of proteins to which DMAHP belongs, only the binding activity of AREC3/Six4 had been investigated. AREC3/Six4 had been identified by virtue of its role as a cell-type specific factor that binds to a 13 bp sequence within the Na^+ K^+ ATPase α 1 subunit gene regulatory element (ARE). DNase I footprinting and methylation interference analyses identified the precise region within the ARE to which AREC3/Six4 bound (Suzuki-Yagawa *et al.*, 1992). The minimal region for sequence specific binding was shown to be encoded by the Six box and homeobox (Kawakami *et al.*, 1996a). The experiments described in this chapter investigated the DNA binding properties of 3 GST-DMAHP recombinant fusion proteins, GST-Six, GST-HD and GST-Six+HD.

5.7.1 Non-denaturing Polyacrylamide Gel Electrophoresis

Initially a standard low-ionic strength polyacrylamide gel and running buffer were utilised. Low-ionic strength buffers generate less heat than high ionic strength buffers and increase the speed of migration by raising the fraction of the current carried by the macromolecules. They also reduce the disruption of ionic bonds which may increase the stabilisation of complexes entering the gel (Kleinschmidt *et al.*, 1991). The Tris/acetate buffer chosen had been successfully used in other DNA/protein binding studies (Chodosh *et al.*, 1986). However, it was not suitable for the recombinant proteins and DNA fragments used in this study as no complexes were detected using this method. Specific complexes formed between phage Mu repressor and its operators could not be detected using a standard TBE buffer but could be detected with a Tris/glycine (pH 9.4) buffer (Alazard *et al.*, 1992; Bednarz *et al.*, 1990). Therefore, experiments were repeated using this buffer and complexes were detected.

5.7.2 Specific Binding of DMAHP to the ARE

Recombinant GST-Six+HD and GST-HD both bound to the ARE but not to the mutant ARE that differed by only a single base pair within the 13 bp AREC3/Six4 (C-G at nucleotide 6) recognition sequence, indicating that this binding was specific. GST-Six did not bind to the ARE, indicating that the Six domain is not directly involved in sequence specific DNA-protein interactions at this DNA site. However, GST-Six was insoluble in the cell lysate and therefore underwent a further stage of

purification to remove inclusion bodies (section 4.3.2c). It is possible that any DNA binding activity was destroyed by this aggressive procedure. When present in conjunction with the homeodomain, the Six domain enhances the affinity of the homeodomain for the ARE and allows two complexes to form, rather than one. It is likely that the upper complex is due to a protein dimer binding to the DNA (see section 5.7.3). Further evidence that the extremely weak binding of GST-HD to the ARE was sequence specific was demonstrated by the addition of a 10 fold molar excess of unlabelled ARE or AREmut to the gel retardation assay. The excess ARE successfully competed for the protein, but an excess of AREmut did not.

An AREC3/Six4 homeodomain GST fusion protein also showed weak affinity for the ARE (Kawakami *et al.*, 1996a). GST fusion proteins containing both the Six domain and the homeodomain of AREC3/Six4 displayed a much stronger affinity for the ARE. Binding of the AREC3/Six4 homeodomain containing fusion protein was not specifically to the 13 bp AREC3/Six4 binding site. An excess of a specific oligonucleotide (C3WT, figure 5.9), which contains the 13 bp sequence but not the entire ARE did not compete for the binding of this fusion protein to a $Na^+ K^+$ ATPase α 1 subunit gene (ATP1A1) *Pvu* II-*Mlu* I fragment (figure 5.9). This indicates that the AREC3/Six4 homeodomain containing fusion protein was binding elsewhere on the ARE. The fact that the DMAHP GST-HD fusion protein complex was not competed by an excess of unlabelled mutant ARE fragment that differed only by a single base pair within the 13 bp AREC3/Six4 binding site, indicates that it was binding to at least the part of the 13 bp sequence where the mutation is. This indicates the possibility that the true binding site of DMAHP is similar but distinct from that of AREC3/Six4. Therefore, it is possible that DMAHP binds to and regulates the transcription of genes which contain a DNA binding site that is similar to the 13 bp sequence in the ARE to which AREC3/Six4 binds. However, GST-Six+HD contains 17 amino acids from the 3' end of the Six domain and this may be sufficient to alter the binding specificity of the homeodomain.

GST fusion proteins encoding the Six domain and the homeodomain of murine Dmahp/Six5 and Six2, but not Six3 have now been shown to bind specifically to the same region of the ARE as AREC3/Six4. Six3 bound to the ATP1A1 *Pvu* II-*Mlu* I fragment, but unlike Six2, AREC3/Six4 and Six5 the binding could not be competed by a 20 fold molar excess of C3WT (Kawakami *et al.*, 1996b).

A binding site has been reported for 3 sine oculis homeodomain containing peptides that were expressed in *E. coli*. The peptides were modified by covalently linking the chemical nuclease 1, 10-phenanthroline-Cu to a unique native cysteine (OP-Cu). The sine oculis -OP peptides were used to locate preferred binding sites in the 3.1 kb

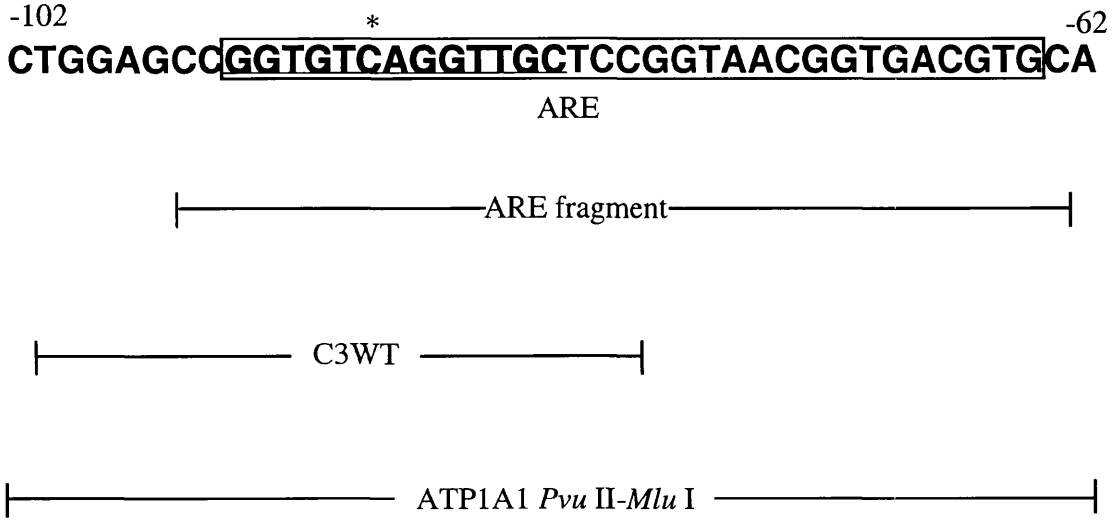

Figure 5.9 Sequence of the ATP1A1 indicating regions used as gel retardation DNA fragments

The sequence of ATP1A1 from -102 to -62 is shown, with the ARE fragment used for this thesis and the fragments described in Kawakami *et al.* (1996a) indicated. The 13 bp sequence to which AREC3 has been reported to bind, is underlined. The asterisk indicates the C that was changed to a G in AREmut. The ARE is defined by a box.

pUC119 plasmid. (C/T)GATAC was identified as a core consensus sequence for the 3 peptides (Hazbun *et al.*, 1997). As this sequence is present 4 times in the related plasmid pBluescript SK(+) which was included in gel retardation assays as a non-specific competitor, there was concern that the recombinant fusion proteins might be binding preferentially to the plasmid rather than to the fragments against which they were being tested. However, when the assays were carried out in the absence of the plasmid, the affinity of the purified GST-Six+HD for the ARE and AREmut did not significantly change. However, the GST-Six+HD in crude cell lysate, bound with relatively low affinity to AREmut (figure 5.3). This indicates that there may be a site or sites in pBluescript SK(+) for which GST-Six+HD has a higher affinity than it has for AREmut. The 3 peptides contained the homeodomain of sine oculis, with the longest containing only 28 amino acids of the Six domain and the shortest containing none of the Six domain. Therefore, these experiments defined a preferred binding site for the homeodomain of sine oculis, without the Six domain. As the experiments with AREC3/Six4 indicated, the preferred *in vitro* DNA binding sites of the Six+homeodomain and the homeodomain are not necessarily the same (Kawakami *et al.*, 1996a).

5.7.3 The Presence of a Cofactor in Bacterial Crude Cell Lysate Enhances the Binding Activity of GST-Six+HD

Unpurified GST-Six+HD showed a stronger affinity for the ARE than affinity purified GST-Six+HD. Control experiments using unpurified recombinant GST indicated that it was the recombinant GST-Six+HD in the cell lysate and not a bacterial protein that was binding to the ARE. Initially it was assumed that GST-Six+HD had been modified or degraded in such a way by the affinity purification process that it had lost most of its binding activity. Western blot analysis indicated that either the concentration of the protein had been drastically reduced by affinity purification and/or the folding of the protein had been altered in such a way as to make to make the antigenic region of the recombinant protein less accessible to the R254 DMAHP antiserum. However, adding recombinant GST expressing cell lysate to affinity purified GST-Six+HD enhanced the binding activity of the protein to levels that were greater than those seen using cell lysate that had expressed GST-Six+HD. Therefore, any changes that occurred during affinity purification were reversible by the addition of cell lysate. The binding activity of GST-HD to the ARE was not enhanced by the presence of cell lysate. Therefore, it is likely that any factor present in the cell lysate that enhanced binding activity requires the Six domain to do so. It is possible that a cofactor binding to the Six domain altered the folding of the fusion protein, thus

allowing R254 DMAHP antiserum access to the antigenic region of GST-Six+HD. However, this does not explain the greater affinity of R254 DMAHP antiserum for unpurified rather than affinity purified GST-HD.

Although many homeodomain proteins bind as monomers *in vitro*, evidence is now emerging that many can and do bind as dimers (section 1.7.3.4). GST-Six+HD but not GST-HD formed 2 complexes with the ARE and the higher molecular weight complex was presumed to be a dimer complex. sine oculis and eyes absent (both *Drosophila* eye development proteins) interact in yeast and *in vitro* through the Six domain of sine oculis and an evolutionarily conserved domain in eyes absent (Pignoni *et al.*, 1997). Therefore, it is likely that the Six domain of DMAHP also acts as a protein binding domain, which *in vitro* allows GST-Six+HD to bind to DNA as a homodimer. To confirm that the upper complex is indeed a dimer, cross-linking experiments could be performed and the products analysed by SDS-PAGE as was done for the POU protein Pit-1 (Ingraham *et al.*, 1990). The Six domain may also be a DNA binding domain and both domains may be required to interact with the DNA to allow high affinity specific DNA binding as is the case with homeodomains containing a paired domain or a POU domain (section 1.7.3). GST-Six may have lost DNA binding activity due to its method of production (section 5.7.2). The presence of the bacterial cell lysate greatly enhanced the formation of the putative dimer complex which suggests that it contains a cofactor that alters the structure of the DNA in such a way as to increase the likelihood of dimer formation. There are many examples of prokaryotic and eukaryotic factors that can bend DNA to enhance the interaction of 2 or more transcription factors (Werner and Burley, 1997). Eukaryotic high mobility-group (HMG) domain proteins bind to the DNA in the minor groove, through a conserved domain, and bend the double helix. There are 2 classes of HMG-domain proteins, one of which contain multiple HMG domains and bind DNA with little or no specificity and the second class contain a single HMG domain and bind to a specific DNA sequence (Love *et al.*, 1995). Similar factors are present in prokaryotic cells and again there are sequence specific DNA binding factors, for example the integration host factor (IHF) and non sequence specific factors, for example HU (Rice *et al.*, 1996). It appears that a cofactor is present in the bacterial crude cell lysate that is sufficiently similar to a natural cofactor present in the human cell to alter the structure of the DNA enough to enhance dimerisation. It may be possible to determine if the cofactor is a small DNA binding protein by dialysing the crude cell lysate to remove small proteins and then testing the lysate to see if it retains its ability to enhance protein-DNA interactions and the formation of the dimer complex.

5.7.4 Solubility of the GST-DMAHP Fusion Proteins

GST-Six+HD and GST-HD were both soluble in the crude cell lysate, whereas GST-Six formed inclusion bodies, possibly due to the greater levels of protein produced (section 4.6.2). However, the 3 fusion proteins were all partially soluble in 50% (v/v) acetone or ethanol and the only one of the 3 that was successfully concentrated by precipitation with organic solvent was GST-HD (section 4.4). The precipitation, however, led to a loss of detectable binding activity. This may have been due to the solvent causing irreversible denaturation of the protein (Schein, 1990).

5.7.5 DMAHP as a Regulator of *DMPK* Transcription

Experiments using GST-HD and GST-Six+HD recombinant proteins in gel retardation assays indicate that the homeodomain of DMAHP does not bind to the GGATTA consensus site as do other homeodomain proteins with a lysine at position 50 of the homeodomain. However, as DMAHP (and other members of the Six subfamily) lack the arginine at position 5 of the homeodomain, which is known to bind to the ATTA core sequence, this is perhaps not surprising. It also appears that DMAHP does not bind to the ARE-like sequence in the promoter of *DMPK* and therefore if DMAHP is controlling DMPK, it is not doing so by binding to either of these putative DNA binding sites.

5.8 Conclusion

The DNA-protein binding activities of 3 GST-DMAHP recombinant proteins have been characterised. GST-Six displayed no detectable DNA-protein binding activity to any of the DNA fragments tested. GST-HD showed weak, but specific binding activity to the 40 bp ARE fragment to which AREC3/Six4 has been reported to bind. GST-Six+HD bound with moderate affinity to the ARE fragment to give 2 complexes. The upper complex was presumed to be formed by a protein dimer and the lower complex by a protein monomer. The binding of affinity purified GST-Six+HD was greatly enhanced by the addition of a bacterial crude cell lysate to the reaction. The experiments suggest that the Six domain of DMAHP is a protein-protein binding domain that enhances the specific DNA binding activity of the homeodomain.

Chapter

6

Whole Genome PCR to Identify Target DNA Binding Sites of DMAHP

6.1 Introduction

The experiments described in the previous chapter indicate that *in vitro*, GST fusion proteins containing the homeodomain of human DMAHP display specific binding activity to the ARE within the promoter region of the Na^+ K^+ ATPase α 1 subunit gene. However, DM is a multisystemic disease and as such, several genes are likely to be implicated in the manifestation of its symptoms. Therefore, it is expected that DMAHP binds to and regulates a number of genes that contribute to the symptoms associated with DM.

To identify other genes that may be controlled by DMAHP, a whole genome PCR based screen was performed using GST-Six+HD bound to a Glutathione Sepharose® 4B affinity matrix (section 2.3.18). Whole genome PCR for the identification of sequences bound by gene regulatory proteins was first used to identify human sequences that bound to Xenopus transcription factor IIIA (Kinzler and Voglestein, 1989). Immune precipitation was used to separate protein bound DNA from unbound DNA. This system was adapted for use with GST fusion proteins bound to a Glutathione agarose matrix, which dispensed with the need for the immunoprecipitation step (Fainsod *et al.*, 1991). GST fused to the chicken *CHox-cad* protein was bound to Glutathione agarose beads and used to identify genomic fragments capable of being bound in a gel retardation assay by extracts from bacteria producing the recombinant fusion protein (Margalit *et al.*, 1993). Target sites for recombinant Sp1 DNA binding protein were selected that closely matched the known Sp1 consensus site using synthetic DNA templates that contain randomised 12 bp oligonucleotides flanked by primer binding sites, in place of genomic DNA (Thiesen and Bach, 1990). Therefore, both genomic DNA and random oligonucleotides have been used, successfully, to identify known binding sequences for the proteins tested. Novel target genes have now been identified for an oestrogen receptor (Inoue *et al.*, 1993) and a thyroid hormone receptor (Caubin *et al.*, 1994) using similar whole genome PCR screens.

Although the use of random oligonucleotides is technically less complicated, it was decided to use genomic DNA, as the main objective of the work was to identify genes to which DMAHP binds rather than just the sequence of the binding site. However, it has been demonstrated that the consensus binding site of the chicken homeodomain protein CDXA, identified by DNase I footprinting of chicken genomic DNA fragments isolated by whole genome PCR, was very similar to the sequence deduced from analysis of random oligonucleotides that bound to the protein (Margalit *et al.*, 1993). Therefore, the work described in this chapter could lead to the identification of

a consensus DNA binding site for DMAHP as well as to the identification of target genes.

GST-Six+HD was chosen as the DMAHP fusion protein for these experiments rather than GST-HD because the Six domain was shown to enhance the binding activity of the homeodomain (section 5.3.2). Sequences isolated in the whole genome screen were tested for their ability to be bound specifically by crude cell lysates of bacteria expressing the GST-Six+HD recombinant protein. The sequences which displayed specific binding activity to GST-Six+HD were analysed further by computer analysis.

6.2 Whole Genome PCR Based Screen

GST Six+HD and GST recombinant proteins were immobilised by binding to Glutathione Sepharose® 4B beads (section 2.3.12.5). The beads and recombinant protein were resuspended in 33 mM Tris (pH 8.0), 80 mM NaCl, 6.7 mM DTT, 0.67 mM PMSF and 20% (v/v) glycerol to give a 50% (v/v) slurry. Aliquots of the recombinant proteins were analysed by SDS-PAGE and the amount of protein estimated by comparison with Coomassie Blue stained BSA standards (results not shown). GST-Six+HD-coated Glutathione Sepharose® 4B beads contained approximately 20 ng of recombinant protein per µl of 50% (v/v) slurry and GST-coated Glutathione Sepharose® 4B beads 60 ng/µl.

6.2.1 Optimisation of the Binding of DNA to GST-Six+HD-coated Sepharose® 4B Beads

To determine the optimum length of time required for the maximum concentration of specific DNA to bind to the GST-Six+HD-coated Glutathione Sepharose® 4B beads, a time course was performed (section 2.3.18.2). 1,000 cps of radiolabelled ARE fragment and vector were mixed with 200 µl of GST-Six+HD (about 4 µg of recombinant protein) or GST (about 12 µg of recombinant protein) bound to the beads. The beads were washed in 1 x DNA-protein binding buffer and the DNA eluted in 1 x DNA-protein binding buffer plus 1 M NaCl. The specific activity (cps) of supernatant and beads at each stage of the experiment was measured. The results are displayed in table 6.1.

	100 µl of supernatant 1	100 µl of beads	1 ml of wash buffer	100 µl of beads after washing	100 µl of eluant	100 µl of beads after elution	%age of total DNA in eluant
GST 15 minutes	300	99	14	22	7	9	2%
GST 30 minutes	198	80	9	8	3	3	1%
GST 60 minutes	215	74	11	8	4	3	1%
Six+HD 15 minutes	142	65	11	13	11	8	5%
Six+HD 30 minutes	99	55	9	13	8	5	5%
Six+HD 60 minutes	120	71	10	14	13	9	7%

Table 6.1 Specific activity (cps) of the ARE at each stage of the binding to and elution from GST and GST-Six+HD-coated Glutathione Sepharose® 4B beads

After a 60 minute binding time the amount of DNA that was eluted from GST-Six+HD (measured as a percentage of the total counts in supernatant and the beads before washing) was seven times greater than that eluted from GST. This indicated that the ARE was binding specifically to the GST-Six+HD and not to any other part of the matrix. 50% of the original specific activity was incorporated in the pBluescript® SK(+), for which GST-Six+HD does not display a strong affinity. Therefore the actual percentage of ARE bound to GST-Six+HD-Glutathione coated Sepharose® 4B beads was between 7 and 14. The experiment also showed that the longer the DNA was allowed to bind to the GST-Six+HD-Glutathione coated Sepharose® 4B beads (up to 60 minutes) the greater the percentage of total DNA that bound. As 60 minutes was the longest binding time tested, it was postulated that a longer binding time would result in more DNA binding. However, as protein is susceptible to degradation at room temperature it was decided that a longer binding time was not feasible. Therefore, the time course experiment was repeated using 500 µl of a 50% (v/v)

slurry of GST-Six+HD-coated Glutathione Sepharose® 4B beads (about 10μg of recombinant protein). This did not increase the amount of DNA bound to and eluted from the protein (results not shown). Therefore, 200 μl of a 50% (v/v) slurry of GST-Six+HD-coated Sepharose® 4B beads (about 4 μg of recombinant protein) and a 60 minute binding time were used in the whole genome PCR based screen.

6.2.2 Optimisation of PCR Conditions

1 μg of high molecular weight human lymphocyte DNA was cleaved to completion with either *Sau*3A I or *Tsp*509 I and PCR linkers were added (section 2.3.18.1). PCR conditions for use in the whole genome screen were optimised using 1 ng of each linkered genomic DNA library as templates and V-NL as the primer (section 2.3.18.4). Initially 1 mM and 2 mM spermidine, 1% and 5% (v/v) DMSO and annealing temperatures of 55°C and 59°C were tried. Unlinkered cleaved genomic DNA was used as a negative control to ensure that the V-NL did not amplify unlinkered DNA. 60% (v/v) of the PCR products were electrophoresed in 1% (w/v) agarose gels (figure 6.1). The greatest amount of product, from both libraries was achieved with no spermidine, 5% (v/v) DMSO and an annealing temperature of 59°C. Optimisation was finalised with the use of 3, 4, 5, 6 and 7% (v/v) DMSO (figure 6.1). 5% (v/v) DMSO and an annealing temperature of 59°C were chosen as the optimum conditions.

6.2.3 Whole Genome PCR

200 μl (about 4 μg of recombinant protein) of a 50% (v/v) slurry of GST-Six+HD-coated Glutathione Sepharose® 4B beads was incubated with 300 ng of linkered genomic DNA (either *Sau*3A I or *Tsp*509 I cleaved). The beads were washed in 1 x DNA-protein binding buffer, bound DNA eluted in 1 x DNA-protein binding buffer with 1 M NaCl added and the eluted DNA amplified by PCR (section 2.3.18.3). After five rounds of binding and PCR no further change to the pattern of PCR products was visible (figure 6.2) and they were subcloned into pGEM®-T Easy vector (section 2.3.18.6). Eighteen clones from each library were sequenced using primers T7 and SP6. Seven independent clones were identified in each of the two libraries and the average length of the fragments selected was 420 bp. One clone contained two plasmids with different sized inserts and therefore would not sequence. Table 6.2 shows the lengths of the selected sequences and number of times each clone was represented.

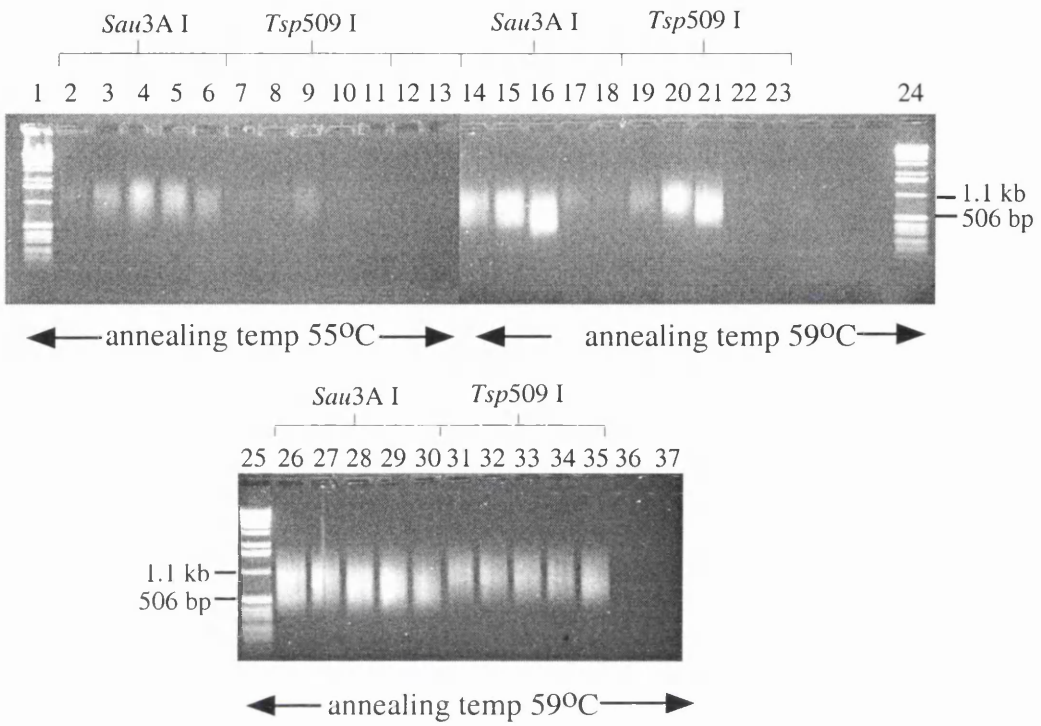

Figure 6.1 Optimisation of PCR for whole genome screen

1 ng of each of the linker libraries, *Sau*3A I and *Tsp*509 I, was amplified using primer V-NL under various PCR conditions. 60% (v/v) of each reaction was electrophoresed in a 1% (w/v) agarose gel. 1% (v/v) DMSO (lanes 3, 8, 15 and 20), 5% (v/v) DMSO (lanes 4, 9, 16, and 21), 1 mM spermidine (lanes 5, 10, 17 and 22) and 2 mM spermidine (lanes 6, 11, 18 and 23) were used with annealing temperatures 55°C and 59°C. 3% (v/v) DMSO (lanes 26 and31) 4% (v/v) DMSO (lanes 27 and 32) 5% (v/v) DMSO (lanes 28, 33, 36 and 37), 6% (v/v) DMSO (lanes 29 and 34) and 7% (v/v) DMSO (lanes 30 and 35) were then used with an annealing temperature of 59°C. 1 ng of unlinkered genomic DNA (lanes 12 and 36) and H_2O (lanes 13 and 37) were used as negative controls. Lanes 1, 24 and 25 contain 1 kb ladder.

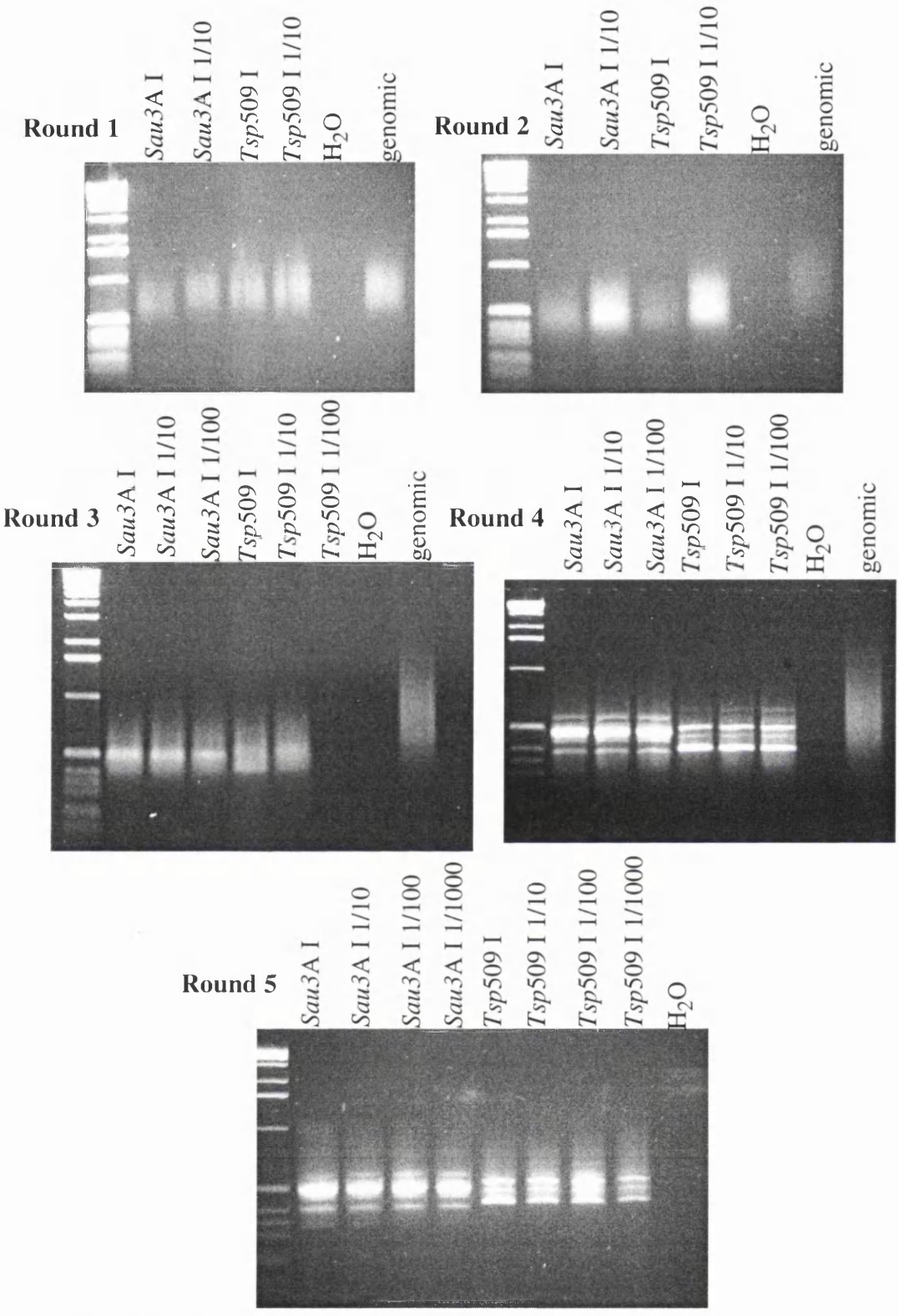

Figure 6.2 Whole genome PCR

Five rounds of binding and PCR were performed. After each round of binding 5 µl of eluate and 1/10 and in later rounds 1/100 and 1/1000 dilutions of eluate were amplified by PCR using primer V-NL. 50% (v/v) of each reaction was electrophoresed in a 1% (w/v) agarose gel. Linkered genomic DNA was used as a positive control for each round of PCR.

Clone	Length of sequence (bp)	Number of times represented
SauA	336	4
SauB	448	5
SauC	524	2
SauD	424	1
SauE	437	3
SauF	464	1
SauG	414	1
TspA	469	8
TspB	359	5
TspC	383	1
TspD	434	1
TspE	388	1
TspF	406	1
TspG	391	1

Table 6.2 Analysis of selected sequences

Interestingly, all sequences contain the expected restriction endonuclease site at only one end of the novel sequence. As the linkers were joined to the digested genomic DNA by the ligation of two halves of a staggered restriction site it was expected that the genomic sequence in each clone would start and finish with such a site, as shown below. **Bold** indicates genomic sequence and a *Sau*3A I restriction site is underlined.

GTCC**GATCNNNNN**GATCGGAC

CAGGCTAG**NNNNNCTAG**CCTG

Two possible explanations for this are:

1) Non-specific internal priming of PCR primer V-NL may have allowed preferential amplification of smaller fragments of DNA during later rounds of PCR.

2) Linkers may have ligated to *Sau*3A I and *Tsp*509 I half sites that have been overdigested by the enzymes. Restriction endonucleases can contain contaminating exonucleases which remove bases from 5' or 3' ends of digested DNA if the reactions

are incubated for longer than the optimum time. For this experiment restriction endonuclease reactions were incubated overnight to ensure that all sites were cut.

In an attempt to determine the cause of the above anomaly, PCR was performed on DNA that did not bind to GST-Six+HD-coated Glutathione Sepharose® 4B beads, during the first round of selection. PCR products were shotgun subcloned into pGEM-T Easy vector. DNA of one clone from each library was sequenced. Both DNA sequences were shown to contain the relevant restriction site at only one end of the sequence.

6.3 Gel Retardation Analysis of the Selected Sequences

To investigate whether DNA recovered from the whole genome PCR selection represented genuine GST-Six+HD binding sites, each independent clone was tested by a gel retardation assay. The sequence containing the selected DNA was released from pGEM®-T Easy by digestion with *Not* I and both vector and insert were radiolabelled (section 2.3.17.2). Gel retardation assays (section 2.3.17) were performed using 10 µl of crude cell lysates from bacteria expressing either the GST-Six+HD recombinant protein or GST diluted 1/10 in 1.5 x binding buffer or 10 µl of 1.5 x binding buffer and 100 cps of each of the 14 selected DNA sequences. At the time these experiments were performed it was not known that affinity purified GST-Six+HD mixed with cell lysate from bacteria expressing recombinant GST showed a higher affinity for the ARE than unpurified GST-Six+HD. Reactions were analysed in 4% (w/v) non-denaturing 1 x Tris/glycine polyacrylamide gels (figure 6.3). GST-Six+HD bound to all 14 clones to give 1 or more retarded complexes. However, 1 of the complexes formed with each of clones SauD, SauE, TspA and TspC also formed with crude cell lysate from bacteria expressing recombinant GST, suggesting that these complexes were formed by the interaction of a bacterial protein or the GST-tag with the DNA. In addition to the complexes formed with recombinant GST containing cell lysates, each of these 4 sequences also formed other complexes only in the presence of cell lysate from bacteria expressing GST-Six+HD. This suggested that these other complexes were formed between GST-Six+HD and the DNA fragment. To test the specificity of the binding of GST-Six+HD to the 14 sequences, gel retardation assays were performed with crude cell lysates from bacteria expressing the GST-Six+HD recombinant protein in the presence of a 100 fold molar excess of unlabelled ARE fragment (figure 6.3). The complexes formed by the binding of GST-Six+HD to 5 of the clones, SauD, TspA, TspC, TspD and TspE were not competed by the presence of excess ARE as were the complexes formed between the GST containing cell lysate and selected DNA. GST-Six+HD showed the strongest affinity

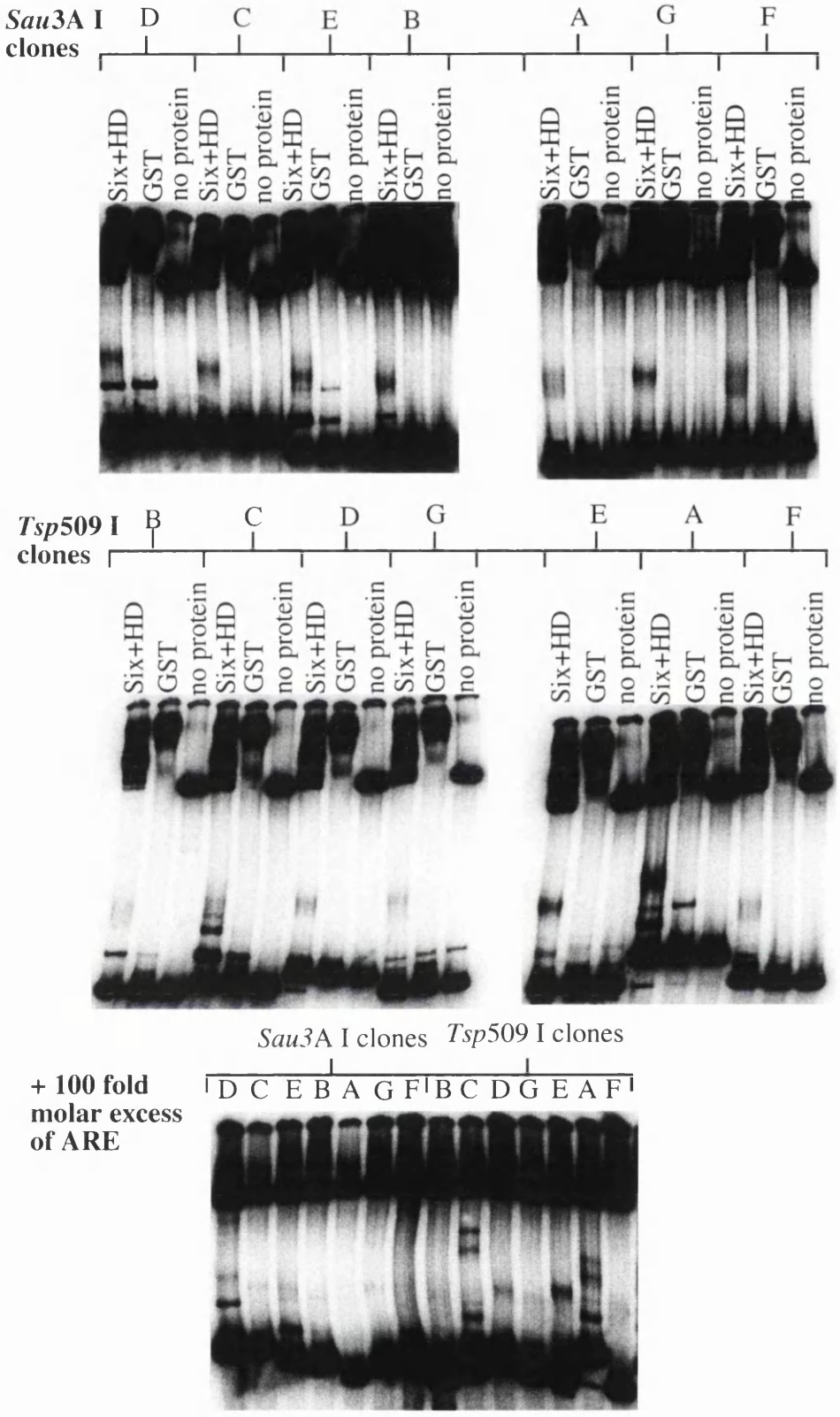

Figure 6.3 Gel retardation analysis of whole genome PCR selected sequences

The selected sequences from each linker library were tested for binding to DMAHP by gel retardation assays with cell lysates from bacteria expressing GST-Six+HD and GST. No protein negative control assays were also performed. Gel retardation assays were then repeated using cell lysate from bacteria expressing GST-Six+HD and a 100 fold molar excess of unlabelled ARE fragment to determine specific binding.

for TspA, which formed 3 complexes with GST-Six+HD and a fourth non GST-Six+HD complex. SauD formed 2 complexes, one of which was with GST-Six+HD. TspC formed 4 complexes of which 3 were with GST-Six+HD and in the presence of excess ARE, 2 further complexes formed. TspD and TspE formed 1 and 2 complexes respectively. The DNA sequences of these 5 clones are shown in figure 6.4.

6.4 Computer Analysis of Selected Sequences

The sequences of SauD, TspA, TspC, TspD and TspE were tested by Bestfit analysis, in both the forward and reverse orientation, to determine a likeness to the AREC3/Six4 minimal sequence as defined in Suzuki-Yagawa *et al.* (1992) (GGNGNCNGGTTGC) (table 6.3).

AREC3SEQ	G	G	N	G	N	C	N	G	G	T	T	G	C
SauDF	G	G	A	G	C	C	T	G	G	**G**	**G**	**A**	**G**
SauDR	G	G	A	G	G	C	T	G	G	**G**	**A**	**T**	C
TspAF	**A**	G	G	G	T	C	A	G	G	T	T	**T**	C
TspAR	G	G	A	**T**	A	**G**	G	G	G	T	T	G	C
TspCF	G	**T**	G	G	T	C	T	G	G	**A**	**A**	**T**	C
TspCR	G	G	T	G	G	C	T	G	G	**G**	T	**T**	C
TspDF	G	G	G	G	A	C	A	G	**C**	**A**	**G**	**C**	C
TspDR	**T**	G	A	G	G	C	A	G	**A**	T	T	**C**	**A**
TspEF	G	G	A	G	C	C	T	G	**T**	T	**C**	**C**	**T**
TspER	G	G	T	G	T	C	C	**A**	G	T	**A**	G	**A**

Table 6.3 Bestfit analysis of the AREC3/Six4 binding site to the selected sequences. The letters in **bold** indicate a deviation from AREC3/Six4 consensus site. F indicates a match with the AREC3/Six4 consensus site in the forward orientation and R a match in the reverse orientation.

The following weak consensus site was identified: GNGNCA/TG. LALIGN analysis (HGMP resource centre) which finds the 10 best local alignments between 2 sequences and will align sequences as short as 10 nucleotides, was utilised. Each sequence was compared with every other sequence in both forward and reverse orientations. Any short sequences present in 2 of the clones were searched for in the other clones in an attempt to identify a strong consensus sequence. None was identified.

SauD

```
   1  GATCCCAGCC TCCTACTTGG GAGGCTGAGG TGGGAGGATT GCTGGAGCCT

  51  GGGGAGGTTG AGACTGCAGT GAGCCATGAT TGCATCACTG CACTGCAGCC

 101  TGGGTGACAC AGCAAGAGCC TGTCTCAAAA GAAAAATAAG ATACCTCAGA

 151  CAAAGCCAGA AAACCAAATA AACAAGGAAG AGTGAGGCAA GGTCTCTGAA

 201  AATGTCCTGT TCTCGGGAGC TATTATGGCA GGAGCCCCCA CCCACGTTGC

 251  TGAATGTTTT ACTTTTCATG AGAGTTCTCA GTTGCACAGT GTGTTAGTCA

 301  GGGTTCTCCA GAGAAGCAAA ACCAACAGGA TGTATACACA GAGAGAAAGG

 351  GAGTTATTAC AAAGCACTGG CTCCCACAAT TACGGAGGCT GGCAAGTCCC

 401  AAGCTCTGCA GGGGGAATTG GCCA
```

TspA

```
   1  AATTGGGTGC ACTTTGGAGG CAAAGCTGGG GCAGAAGAGT GGAATAAGGA

  51  TGGACTCATG TGGGAACAGT GGGCCTGCTG TTCCCTGGAG CCCCACAGGG

 101  TCAGGTTTCC CATCTCTGAC TGCAACCCCT ATCCCACTCC CACAGGAAGG

 151  TGTCATTTCT GAGAGACAGC CAGACTTACG TGTGTATTAG TCTGTATTCA

 201  CACTGCTGAT AAGGACACCC GAGACTGGGA AGAAATAGAG GTTTAGTTGG

 251  AGTTACAGTT CCACGTGGCT GGGGAGGCCT CAGAATCATG GTGGGAGGCA

 301  AAAGTACTTC TTACATCATG GCGGCAATAG AAAATGAAGA AGCACAAGTG

 351  GAAAACCCCT GGTAAACCCA TCAGATCTCG TGAGACTTAT TCACTATCAC

 401  GAGAATAGCA TGGGAGAGAC TGGCCCCCGT GATTCAGTTA CCTCCCCCTG

 451  CGTCCCTCCC ACATGTGGG
```

TspC

```
  1  AATTCTATGT GAGGGACAAA CACTCAGAAC CCATCAGCAG TGTTCTGGAA
 51  TCCTACATGA TGGACAAACA CTCTGAACCC AGCAGCAGTG TTCTGGCATC
101  CTATGTGGGG GACAAACACT CAGAACCCAG CCACCGTGTT CTGGAATCCT
151  ATCTGTCTGA TAAATATTCA GACACTCATA GAAATGTTCT GGAATCCTAA
201  GTGATGGAGA AACACTCAGA AATCAGCAGC AGTCCTCTGG AATCCTTTGT
251  GAAGGATAAA CAAAGAGAAC CCACCAGCAG TGTTCTGGAA TCTTTCTGAG
301  TGAAAAACAT TCAGAACCTA GTAGCAGTGG TCTGGAATCC AATGTGAGGG
351  AGAAACACTC AGAACCCAGC AGCAGTGTGC TGG
```

TspD

```
  1  CCACTGGGGA CAGCAGCCTA CTTACAGAAG CGAAGCCAAC TGCAGCTGGG
 51  ATAAAGTGAT AATAGATAGG ACTGACAAGG AGGCGTGGCC TTCCATCACA
101  GGAACAGAGA CTGAATCTGC CTCAGAATGT ACTACAGACA CTGACTCTGC
151  CTCCAACTGT GGCTCAGAGA ACAGTAGCAT GGCTACAGGG AGTGCCCAGG
201  GCAACTTCAC TGGACATACC AAGAAGACAA ATGGCAATAA TGGCACCAAT
251  GGCGCACTCG TCCAAAGCCC TTCTAATCAG AGTGCCCTTG GAGCAGGGGG
301  AGCGAACAGT AATGGAAGTG CGGCCAGAGT GTGGGGTGTA GCCACAGGCT
351  CCAGCTCTGG CCTGGCTCAC TGCTCTGTCA GTGGTGGGGA TGGAAAAATG
401  GACACTATGA TTGGAGATGG GAGAAGTCAG AATT
```

TspE

```
  1  CCAGCCTCAT GGTGTTGAGG GAGGGGCTGG GGGAAGCGGG AGCCTGTTCC
 51  TAGGAGACAC TGTCTACTGG ACACCTCATC CTAGAAGGCT TCCCTGAGTT
101  CTGACACTGT TCTAGGAGTA GGAATCATTA TCCCACGCTG TGGAGAAGGA
151  AACCAAGGCC GGGTAAAGTA ATGTAGGGCC CTTAAAGTTG CAGAGCCAGG
201  AATAAAGTCC GGATTGGGAC TCAGGTCTGT GGGCAATCCA AGGCCGTCCC
251  CTTTACACTT CAAATGCTCT TTCCCCCTCC GGAAGCCCTC GCGTCTTCAT
301  CCCTACCCCA CCTCTTGTTC CCCAAGCGTG GCTAGGGCTA GGGCTCCAGG
351  GCTACACCAA GCACCCTTCG GTCTTCCCGG GAAGAATT
```

Figure 6.4 Sequences that bind specifically to GST-Six+HD
Sequences identified by Bestfit analysis are underlined.

The sequences of each selected clone were compared to the GenBank, EMBL, DDBJ and PDB databases using the BLASTN program at NCBI. TspE showed 92% identity to a human dopamine D5 receptor gene, 5' flanking and promoter region. The 388 bp TspE sequence maps from -965 to -578 of this sequence (accession number U21164) and starts 1161 bp 3' of the transcription initiation site (Beischlag *et al.*, 1995). TspC contains non-alphoid repetitive sequence, but matches no specific region of the genome. SauD, TspA and TspD showed no significant matches to any sequences in the databases.

6.5 DNase I Footprinting of TspA

Confirmation of the GST-Six+HD binding site(s) within the selected sequences requires DNase I footprinting. It was decided to footprint sequence TspA as this bound most strongly to GST-Six+HD. To prepare a DNA fragment for use in a DNase I footprinting experiment TspA was digested with *Nco* I and *Nsi* I to release a fragment containing the selected sequence (figure 6.5). This fragment was gel purified from a 1% (w/v) agarose gel using a Qiagen gel extraction DNA purification kit. As *Nsi* I leaves no overhanging Gs the fragment was radiolabelled at the *Nco* I end only, using [α-^{32}P] dCTP (section 2.3.19.1). To create size markers for electrophoresing alongside the DNase I reactions, restriction endonuclease sites were identified within the sequence. *Alu* I, *Dde* I, *Bst*X I and *Sau*3A I cut 25, 162, 264 and 377 bp respectively from the start of the selected sequence closest to the radiolabelled end of the fragment released from pGEM®-T Easy (figure 6.5).

DNase I footprinting was performed using 10 µl of crude cell lysate from bacteria expressing either recombinant GST-Six+HD or recombinant GST diluted 1/10 or 1/20 with 1.5 x binding buffer or 10 µl of 1.5 x binding buffer and 200 cps of radiolabelled TspA (section 2.3.19.2). No discrete footprint could be distinguished, but evidence of GST-Six+HD binding to the sequence was indicated by changes in the DNase I restriction pattern in the reactions containing GST-Six+HD when compared to reactions containing GST or no protein (figure 6.6). The decrease in intensity of bands was noted in the regions identified by Bestfit analysis as being those most closely resembling the AREC3/Six4 binding site. The assay indicated the presence of a substance in the crude cell lysate that enhanced the activity of the DNase I. Although the pattern of digestion was similar in all lanes, digestion was increased in lanes containing cell lysate.

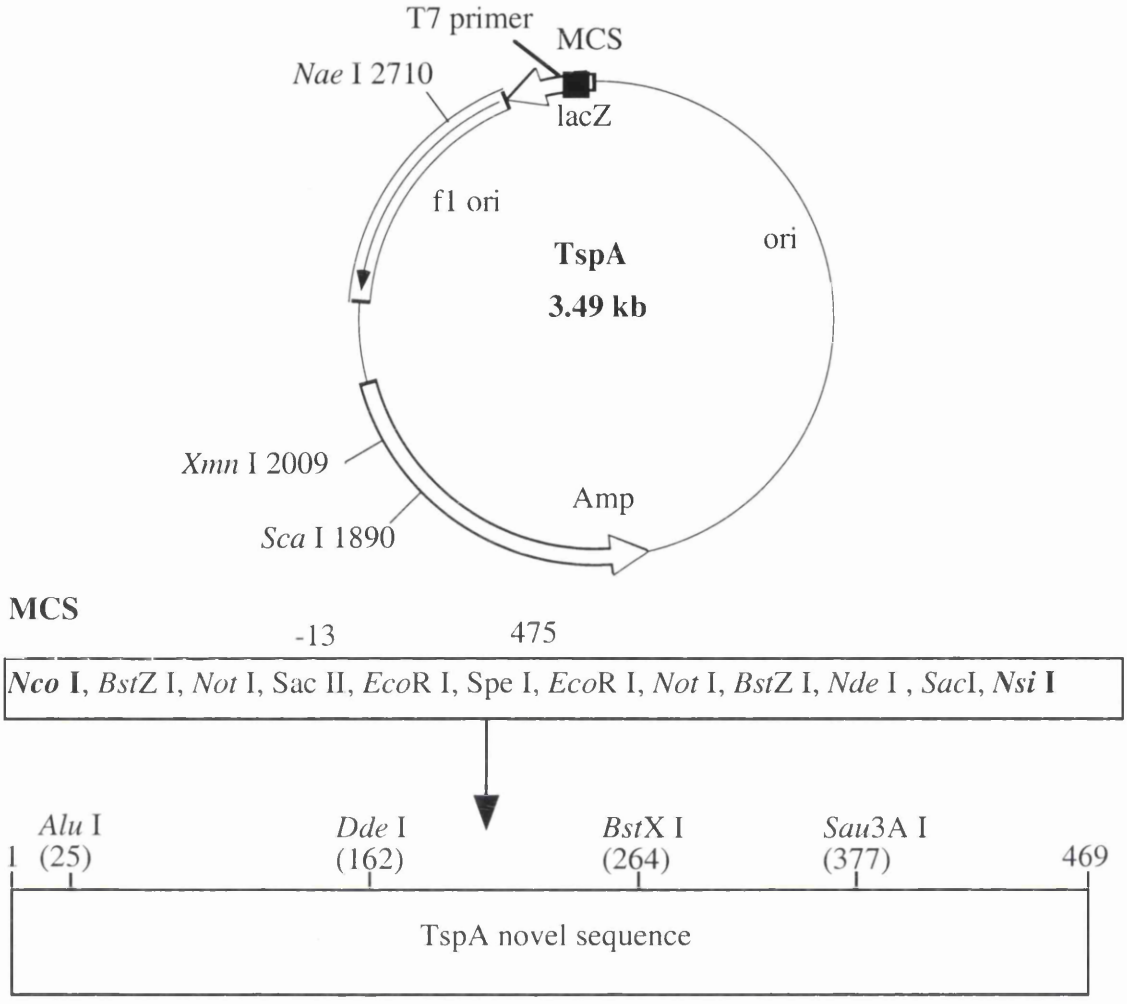

Figure 6.5 Structure of TspA

A schematic diagram showing the structure of clone TspA and the restriction sites that were used to create size markers for DNase I footprinting of the clone: *Alu* I, *Dde* I, *Bst*X I, *Sau*3A I.

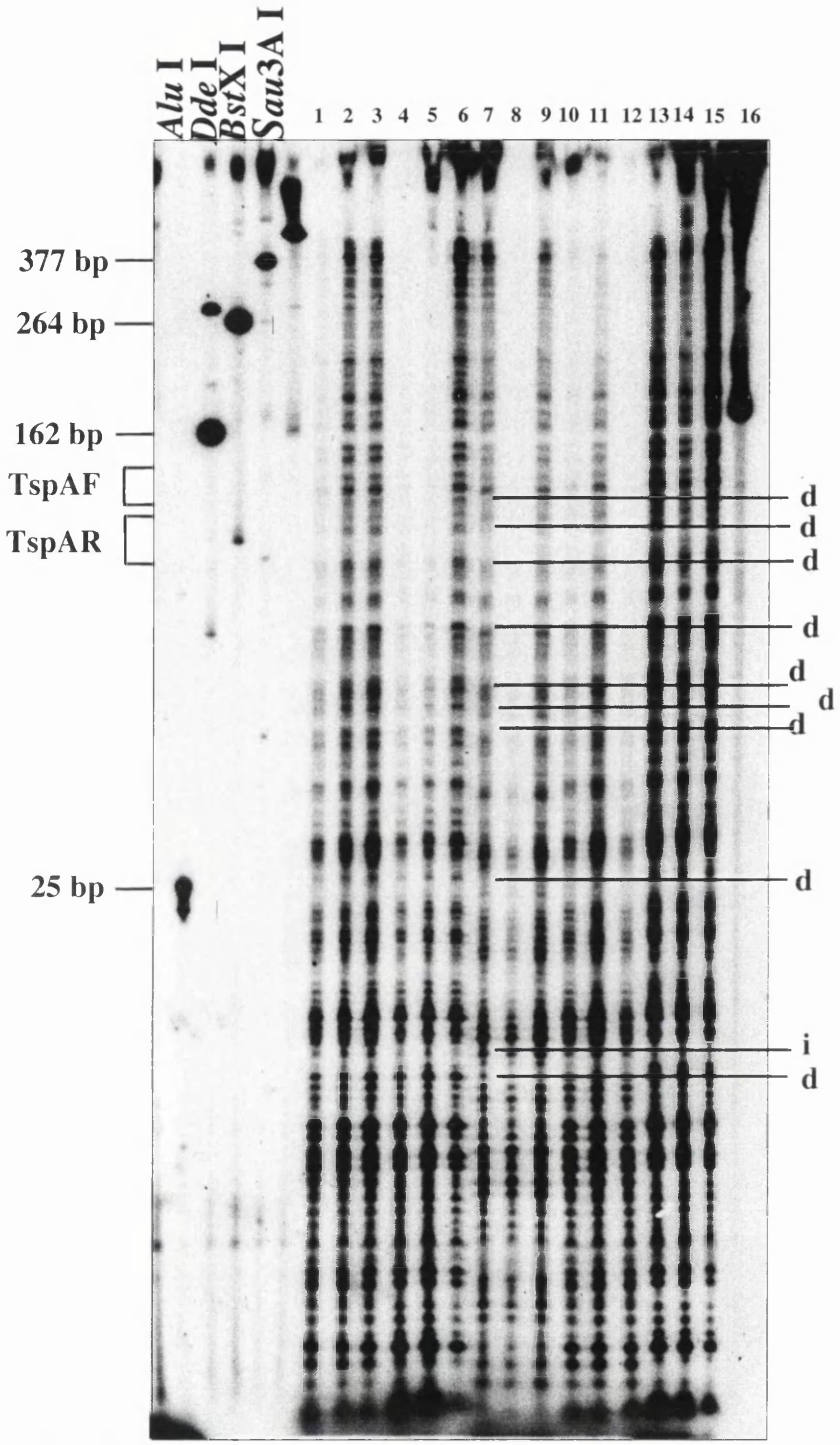

Figure 6.6 DNase I footprint of TspA

TspA was analysed by DNase I footprinting with GST-Six+HD. The lanes are as follows: 1/10 dilution of GST lysate with 12.5 ng, 6.25 ng and 3.13 ng of DNase I (lanes 1, 2 and 3 respectively), 1/20 dilution of GST lysate with 12.5 ng, 6.25 ng and 3.13 ng of DNase I (lanes 4, 5 and 6 respectively), 1/10 dilution of GST-Six+HD lysate with 12.5 ng, 6.25 ng and 3.13 ng of DNase I (lanes 7, 8 and 9 respectively), 1/20 dilution of GST-Six+HD lysate with 12.5 ng, 6.25 ng and 3.13 ng of DNase I (lanes 10, 11 and 12 respectively), no protein with 12.5 ng, 6.25 ng and 3.13 ng of DNase I (lanes 13, 14 and 15 respectively), no DNase I control (lane 16). d=bands that decreased in intensity in the presence of 1/10 dilution GST-Six+HD and 12.5 ng of DNase I. i = bands that increased in intensity. The positions of TspAF and TspAR are indicated.

6.6 Discussion

The principal objective of the whole genome PCR based screen was to identify human genomic sequences to which recombinant GST-Six+HD binds. Sequences identified by the screen can be utilised to identify genes that are controlled by DMAHP. To date no *in vivo* targets of DMAHP have been identified and therefore this screen was the first stage in the identification of what is potentially a significant number of genes controlled by DMAHP.

6.6.1 Analysis of Isolated Sequences

Sequencing of 35 clones indicated that 14 independent sequences with an average length of 420 bp had been identified. The anomaly of the sequences containing the expected restriction site at only one end was investigated. The fact that after only one round of PCR, sequences contained a restriction site at just one end suggests that it was not caused by internal priming leading to the preferential amplification of smaller fragments. It also supports the argument that the assay was not biased for a small number of fragments containing a restriction site at only one end. Therefore, the most likely explanation is that the restriction endonuclease digestion reaction to generate fragments of genomic DNA for linker ligation was incubated for too long and that nucleotides were removed from the ends of the DNA fragments through exonuclease activity. This will have lead to the inefficient ligation of linker DNA. Linkers at DNA ends with incomplete sequence can only partially ligate and single strand synthesis from a correctly ligated linker would possibly have been required before priming from the partially ligated linker could occur. Therefore, a complete overhang would be required at one end of each fragment for it to be successfully amplified.

The 14 independent sequences were tested by gel retardation assays and GST-Six+HD showed specific binding activity to 5 of them. Four of the sequences formed multiple complexes with the recombinant protein. The higher molecular weight complexes may have been caused by GST-Six+HD binding to the DNA fragment as a homodimer (section 5.7.2). Alternatively, the sequences contained multiple binding sites and the higher molecular weight complexes contained multiple molecules of GST-Six+HD per molecule of DNA. TspC formed 2 extra complexes when a 100 fold molar excess of ARE was added to the binding reaction. This may have been an experimental artefact, but it is possible that TspC contains a DNA binding site for the Six domain of DMAHP and when the ARE was added to the reaction this bound to the

homeodomain and a supercomplex formed, that consisted of GST-Six+HD, TspC and the ARE fragment.

Bestfit analysis of the AREC3/Six4 consensus site to the 5 selected sequences, failed to identify an exact match, although putative sites in each sequence and a weak consensus sequence were identified. LALIGN analysis did not identify a consensus site. However, LALIGN can only analyse 2 sequences at a time and so only strong consensus sites could be identified this way. An algorithm which assumes that the binding site is a contiguous region of DNA whose essential features reside within the individual bases of a sequence can be used to identify consensus patterns in several unaligned DNA sequences (Hertz *et al.*, 1990). However, this algorithm has only been tested on sequences up to 200 bp in length and is not suitable for sequences of the length identified above. Therefore, confirmation of GST-Six+HD binding sites within the sequences identified by the whole genome screen requires DNase I footprinting to be performed on each sequence.

TspE was the only sequence identified that showed high similarity to a known gene, the dopamine D5 receptor gene. Encouragingly, the sequence is from the 5' UTR of the gene which is where transcription factors are known to bind. The sequence is situated between the transcription and translation start sites and may therefore interfere with transcription by binding to the DNA. Alternatively, DMAHP may bind to the RNA at this position and thus act as a translational repressor. A *Drosophila* homeodomain protein, bicoid, which binds DNA and transcriptionally activates different target genes has also been shown to repress the translation of *caudal* by binding to the 3' UTR of the mRNA (Dubnau and Struhl, 1996). Dopamine receptors belong to a superfamily of receptors that exert their biological effects through guanine nucleotide-binding (G) proteins and are the primary targets for genes used to treat psychomotor diseases like Parkinson's disease and schizophrenia. Two main dopamine subtypes have been identified, D1 and D2. D5 belongs to the D1 subtype that stimulate adenylyl cyclase activity. The distribution of D5 receptor gene transcripts are neuron-specific (Sunahara *et al.*, 1991). Two dopamine D5 receptor pseudogenes have been identified (Nguyen *et al.*, 1991) that share 92% and 91% identity to the dopamine D5 receptor gene. It is possible that the sequence identified by the whole genome PCR screen is the 5'UTR of one of these pseudogenes. However, the 5' UTR's of these sequences are not in the databases and therefore at present this cannot be confirmed. Even if the sequence identified is a pseudogene it is likely to contain the same regulatory elements as the dopamine D5 receptor gene and therefore would not eliminate it as a candidate for regulation by DMAHP.

It has been reported that the dopamine receptors of pituitary prolactin cells might be impaired in DM patients (Sakuma *et al.*, 1988). DM patients suffer from psychological problems (Harper, 1989) which may be due to defects in their dopamine receptors. Therefore the dopamine D5 gene is a good candidate for being causative in DM. Further studies into the levels of D5 receptor mRNA and protein in DM patients and normal controls need to be performed to confirm any role in the pathology of DM.

The other 4 sequences identified by the whole genome PCR screen showed no significant matches to any sequences in the databases. Therefore, these sequences need to be characterised further (section 7.5). The Na^+ K^+ ATPase α 1 subunit gene was not identified by this screen, neither was a sequence identical to the AREC3/Six4 minimal sequence as defined in Suzuki-Yagawa *et al.* (1992). This suggests that DMAHP has a greater affinity for some other sequence, possibly related to the AREC3/Six4 binding site, than for the AREC3/Six4 binding site itself as does the fact that 100 x excess of ARE does not compete for the binding of GST-Six+HD to 5 of the sites selected by this assay.

6.6.2 DNase I Footprinting of TspA

An initial DNase I footprinting analysis of TspA, the sequence to which GST-Six+HD bound most strongly, displayed no discrete footprint. However, the activity of DNase I at several regions of the sequence was either enhanced or repressed indicating that although the sequence did not contain a single site for which GST-Six+HD has a strong affinity, it may contain several weaker sites. This would explain the multiple complexes formed in the gel retardation assay.

6.6.3 Conclusion

A whole genome PCR based screen of the human genome, for target DNA binding sites of DMAHP, identified 5 sequences. One of the sequences is contained within the 5' flanking region of the dopamine D5 receptor gene.

Chapter 7

Final Discussion and Future Directions

7.1 Introduction

The initial aim of the work carried out for this thesis was to identify a full length *DMAHP* cDNA that could be subcloned whole and in fractions into a bacterial expression vector for the production of recombinant DMAHP proteins. The failure to obtain such a cDNA was overcome by the use of a genomic cosmid and a RT-PCR product. From these, 3 recombinant proteins were successfully prepared and used to characterise the binding properties of DMAHP, using a known binding site of the related protein AREC3/Six4, and putative binding sites present in *DMPK*. The final objective involved the use of a whole genome PCR based screen to identify genomic regions to which DMAHP binds, as an initial stage in the identification of genes controlled by this transcription factor.

7.2 Expression Levels of *DMAHP* and *DMPK*

Changes in the levels of expression of both *DMAHP* and *DMPK* have been implicated in causing the symptoms associated with DM. Early reports on the levels of *DMPK* expression in DM patients were contradictory (section 1.5.2). However, more recent evidence suggests that equal levels of unprocessed pre-mRNA are produced by both normal and expanded alleles, but that transcripts from the DM allele are retained in the cytoplasm, leading to a reduction in processed mRNA levels from this allele relative to the normal allele. As the size of the $(CTG)_n$ expansion increases, the levels of processed mRNA from the DM allele decreases (Hamshere *et al.*, 1997; Krahe *et al.*, 1995a; Taneja *et al.*, 1995; Wang *et al.*, 1995).

In an attempt to determine the function of *DMPK* and to understand how it may be involved in DM, 2 groups have generated *Dmpk* deficient homozygous null mice (Jansen *et al.*, 1996; Reddy *et al.*, 1996). To nearly all extents they are indistinguishable from wild type mice, displaying few symptoms similar to the DM phenotype. A very slight alteration in the ratio of small to large fibres in skeletal muscle was the only apparent abnormality (Jansen *et al.*, 1996). Interestingly, altered calcium homeostasis in *Dmpk* deficient cultured myocytes was observed, suggesting possible defects in conduction or the actual contractile mechanism. A late-onset, progressive skeletal myopathy was observed by Reddy *et al.* (1996), similar to that seen in DM. Again, variation in fibre size was observed in these mice along with increased fibre degeneration and fibrosis. Therefore, *DMPK* may be necessary for the maintenance of skeletal muscle structure and function. However, the minimal phenotypes detected, were only present in homozygous *Dmpk* null mice and not heterozygotes. As DM is dominant, these models do not accurately simulate the disease. In

mice overexpressing *DMPK*, due to insertion of multiple copies of a *DMPK* transgene, no phenotype suggestive of DM was observed (Jansen *et al.*, 1996). Patchy cardiac hypertrophy was seen, however, this is not the same as the fatty infiltration and fibrosis of the cardiac conduction system seen in DM patients.

These studies appear to have ruled out haploinsufficiency of *DMPK* or overexpression of this gene as the main cause of DM and emphasise the need to investigate neighbouring genes and the effect of the $(CTG)_n$ expansion in general. An animal model where the $(CTG)_n$ expansion itself is replicated might yield more information particularly regarding its effects on neighbouring genes, for example *DMAHP*, the expression of which has now been demonstrated to be reduced in DM patients (section 1.6.2).

DMAHP and *DMPK* have been shown by RT-PCR to be expressed in a similar wide range of tissues (Boucher *et al.*, 1995; Heath *et al.*, 1997) and it was hypothesised that DMAHP is a transcription factor involved in the control of several genes involved in the pathology of DM, including *DMPK*. Therefore, putative DMAHP DNA binding sites were searched for in the promoter of *DMPK* and 2 were identified. Binding to these sites was tested by gel retardation assays and DMAHP showed no affinity for either of them. Therefore, if DMAHP is a transcription factor involved in the regulation of *DMPK* it is doing so by binding to some other site and it is important that any such site be identified. *Cis* elements that may activate expression of *DMPK* preferentially in myoblasts over fibroblasts have been reported (Storbeck *et al.*, 1998). It was demonstrated that *DMPK* transcription is at least partially regulated by a 51 bp MyoD-responsive element located in the first intron of the gene and that this cooperated with *DMPK* promoter elements one of which is an Sp1 binding site. The experiments implied that distinct *cis* elements located elsewhere in the gene are involved in the regulation of *DMPK* transcription in other tissues. Only 2 putative sites in *DMPK* were tested in this project and therefore the promoter region and intron 1 could be divided into fragments and each fragment tested for its ability to bind DMAHP. *DMPK* was not identified in the whole genome screen, but this may have been due to the binding site being located within a genomic fragment that was too long to be efficiently amplified by the PCR or DMAHP may only bind to *DMPK* in the presence of a cofactor. Alternatively, DMAHP may not directly interact with *DMPK* but they may be involved in pathways in which there are intermediate genes (section 7.3).

7.3 Current Knowledge of the Expression of the *Six* Subfamily of Homeobox Genes and their Implication in DM

Since the initiation of this project several papers have been published reporting expression data for members of the *Six* subfamily of genes in several species. The following provides a summary of the recent data and indicates the possible involvement of the *Six* subfamily in the pathology of DM:

Human *DMAHP* has been detected by *in situ* hybridisation in the lens, cornea, retina and ciliary bodies of adult eye (Winchester *et al.*, in press) all of which are sites of pathological changes seen in DM patients. *In situ* hybridisation failed to detect *DMAHP* in any other tissue. However, the more sensitive method of RT-PCR indicates that *DMAHP/Six5* is expressed in a wide range of adult human and mouse tissues and mouse embryonic head, limbs, liver, kidney and heart (Boucher *et al.*, 1995; Heath *et al.*, 1997) and transgenic reporter studies using a 4.3 kb murine *Six5* promoter fragment fused in frame to the *lacZ* reporter gene have also detected expression in a range of embryonic tissues (Heath *et al.*, 1997). *Six5* transcripts have also been detected by *in situ* hybridisation specifically in the inner and outer nuclear layers, ganglion cell layer and pigment epithelium of mouse adult retina (Kawakami *et al.*, 1996b) and the protein by super-shift assays in rat adult lung, liver and kidney and embryonic lung (Ohto *et al.*, 1998). These data suggest that *DMAHP/Six5* is an adult type *Six* gene isoform and the levels of *DMAHP/Six5* expression are relatively low in all tissues except the adult eye, thus making *DMAHP* a particularly strong candidate for being involved in the cataract development and retinal degeneration common to the majority of DM patients. From the results of these experiments it would be advisable for future cDNA library screenings to obtain a *DMAHP* cDNA to concentrate on adult eye tissue cDNA libraries. However, there is a recent report of *Six5* expression being detected by northern blotting in adult mouse heart, lung, kidney, liver, skeletal muscle, spleen and testis and embryonic heart and skeletal muscle (Murakami *et al.*, 1998). Therefore, it is likely that changes in *DMAHP* expression are also responsible for other DM symptoms, for example cardiac defects and myotonia.

AREC3/Six4 mRNA was detected in adult murine retina and protein was detected in the central and peripheral nervous systems from E9.5-14.5 and embryonic lung (Kawakami *et al.*, 1996b; Ohto *et al.*, 1998). Chicken and zebrafish *six4* homologues have now been identified and shown to be expressed in the head mesoderm of the zebrafish (Bovolenta *et al.*, 1996; Seo *et al.*, 1998c). Murine *Six2* mRNA was also detected in the adult retina (Kawakami *et al.*, 1996b) and the protein was observed by immunohistochemistry in the mesenchyme from E8.5-E13.5 (Ohto *et al.*, 1998).

Human *SIX2* transcripts have recently been detected in foetal eyes, brain, liver, heart and limb and adult ovary, heart and skeletal muscle (Catherine Winchester, personal communication). Murine *Six3* is also expressed in adult retina (Kawakami *et al.*, 1996b). A six3 chicken homologue has been identified (Bovolenta *et al.*, 1996) that is expressed in the most anterior portion of the neural plate, structures derived from the anterior neural plate, the optic vesicle, the rostroventral forebrain and later the prospective neural retina and lens (Bovolenta *et al.*, 1998). Three *Six3* zebra fish homologues (*zsix3*, *zsix6* and *zsix7*) have also been identified, with expression detected in optic vesicles and the rostral forebrain (Kobayashi, 1998; Seo *et al.*, 1998a; Seo *et al.*, 1998b). Overexpression of *zsix3* induces rostral forebrain enlargement, enhanced expression of *pax2* in the optic stalk and leads to a general disorganisation of the brain (Kobayashi, 1998). A medaka *Six3* homologue is expressed at late gastrula/early neural stages in the anterior shield region and then in the developing diencephalon, optic vesicle and lens placode (Loosli *et al.*, 1997). A human Six1 homologue (SIX1) has been cloned and transcripts identified in adult skeletal muscle (Boucher *et al.*, 1996).

A sixth human *Six* subfamily gene has recently been identified (*OPTX2*) along with its chicken, mouse (*Optx2*) and *Drosophila* (*optix*) homologues (Toy *et al.*, 1998). Chicken *Optx2* transcripts have been detected in embryonic retina, optic vesicles, early lens placode and corneal epithelium. *Drosophila optix* is a closer *Drosophila* homologue of mammalian *Six3* than *sine oculis* and is expressed during early development of the fly head and eye primordium. A third *Drosophila* member of the family that appears to be a closer relative to *DMAHP/Six5* than *sine oculis* is currently being characterised (Graham Hamilton, personal communication).

All of these recent reports describe the expression of the *Six* subfamily of genes in an overlapping wide range of tissues. DMAHP/Six5 has been shown to bind *in vitro* to the same DNA sequence (the ARE) as AREC3/Six4 and Six2 (this thesis and Kawakami *et al.*, 1996b) which suggests that there may be cooperative and/or competitive interactions between members of the Six subfamily. Na^+, K^+ ATPase is the enzyme responsible for maintaining the Na^+ and K^+ gradients across the cell membrane, after the propagation of an action potential, which may be involved in the myotonia associated with DM. A recent study demonstrated that members of Six subfamily bind to the MEF3 site (TCAGGTT) in the promoter of myogenin, a protein that is required for myoblast fusion (Spitz *et al.*, 1998). The MEF3 site matches bases 5-11 of the 13 bp sequence within the ARE to which DMAHP/Six5, AREC3/Six4 and Six2 bind and it is present in clone TspA described in this thesis (chapter 6). However, it is also present in ARE-like to which GST-Six+HD does not bind (chapter 5) and the whole genome PCR-based screen did not identify Myf-4 (the human

homologue of myogenin). Six1 and AREC3/Six4 were identified as being the proteins responsible for forming 2 muscle specific complexes with MEF3 in a gel retardation assay using nuclear extracts from various adult tissues. A third ubiquitous complex was formed by an unidentified protein, possibly Dmahp/Six5 based on its expression pattern. It is possible that members of the Six subfamily are partially responsible for the various stages of myogenesis. *Six1* is expressed in the somites of mouse embryos prior to myogenin induction (Oliver *et al.*, 1995b) and therefore may control early activation whereas other members of the subfamily may act later in development. A downregulation of *DMAHP* could therefore lead to the muscle atrophy associated with classical DM or the delay of muscle maturation in congenital DM patients.

There is now strong evidence for the involvement of both *DMAHP* and *DMPK* in muscle development and it is possible that they are both factors are active in the same pathway. DMAHP/Six5 is a candidate for the regulation of Myf-4/myogenin which is also known to be regulated by MyoD (reviewed in Weintraub, 1993). *Dmpk* was shown to be expressed in myogenic cell lines and to be upregulated in cells transfected with a *MyoD* expression vector (Sabourin *et al.*, 1997; Storbeck *et al.*, 1998). Upregulation of *DMPK* was also observed in normal human and DM fibroblasts transfected with MyoD (Otten andTapscott, 1995). Therefore, it is likely that MyoD also regulates *DMPK*.

With the exception of *Six1* all members of the Six subfamily of genes have been shown to be expressed in eye. Therefore together the expression and DNA-protein binding data suggest a possible role for the *Six* subfamily in the eye pathology, muscle atrophy and myotonia associated with DM. It would be interesting to investigate the expression of *SIX1*, *SIX2*, *SIX3*, *AREC3/SIX4* and *OPTX2* and the genes which they control, in DM patients and normal controls. If the Six subfamily proteins are involved in cooperative and/or competitive interactions then the downregulation of *DMAHP* expression could lead to an alteration in the expression levels of genes controlled by other members of the subfamily.

To date only one other member of the *Six* subfamily has been linked to a human disease. *SIX3* has been localised to a region on chromosome 2 known to be involved in holoprosencephaly (HPE), *HPE2*. Three missense mutations in the homeodomain of unrelated patients with sporadic HPE have been identified. One mutation changes arginine to proline at position 52 of the homeodomain. This amino acid is conserved between species and all *Six* subfamily members. The second mutation changes leucine to valine at amino acid 21 of the homeodomain. This amino acid is conserved between human *SIX3* and mouse *Six3* and *Optx2* and *Drosophila optix* but not the other *Six* genes. The final mutation changes valine to alanine at position 45 of the

homeodomain. This amino acid is conserved between species and all *Six* subfamily members. (Wallis *et al.*, 1998). These mutations demonstrate the importance of these 3 amino acids for the correct functioning of SIX3. As they are all in the homeodomain it is likely that they affect the DNA binding specificity of the protein. Therefore, they are potential targets for future *in vitro* mutagenesis experiments to produce either dominant gain of function or dominant negative Six3 mutants to determine the normal role of the *Six3* gene.

7.4 The Functional Domains of DMAHP

The DNA binding activities of the Six domain and the homeodomain were investigated by gel retardation assays. GST-Six+HD displayed a greater affinity for the ARE than did GST-HD and no DNA binding properties of GST-Six were identified. GST-Six+HD bound to the ARE as a dimer and it was suggested that the Six domain may act as a protein-protein binding domain as it does in so. *Drosophila* so and eya were shown by genetic and molecular studies to regulate multiple steps in eye development and to have identical mutant phenotypes. Therefore, the yeast two-hybrid system was used to test the interaction of these proteins. The yeast two-hybrid system (Fields and Sternglanz, 1994) works by the interaction of a pair of proteins bringing a transactivation domain into close proximity with a DNA binding site that regulates the expression of an adjacent reporter gene. Candidate proteins can be tested for their ability to bind the target protein and novel proteins can be identified by the use of libraries constructed from total cDNA from either a whole organism or a particular tissue. Deletion studies localised the interacting regions to the Six domain of so and an evolutionarily conserved domain of eya, the eya domain. *In vitro* binding of radiolabelled so and eya to either GST fused to a fragment of so containing the Six and homeodomains (GST-Six+HD) or the eya domain (GST-Eya), immobilised on glutathione agarose beads, confirmed direct binding between so and eya (Pignoni *et al.*, 1997). The identification of proteins with which DMAHP interacts will be an important step in determining the pathways in which DMAHP is involved. Three murine *Eya* genes have been identified and the Eya protein binding domain is highly conserved in these proteins. Members of the mammalian *Eya* family are expressed in the developing eye and somites (Abdelhak *et al.*, 1997; Xu *et al.*, 1997a; Xu *et al.*, 1997b; Zimmerman *et al.*, 1997) and are therefore strong candidates for interacting with DMAHP and other members of the Six subfamily. Yeast two-hybrid and *in vitro* binding experiments could be performed using DMAHP/Six5 and the other mammalian Six proteins (Six1, Six2, Six3, AREC3/Six4 and Optx2) with mammalian

Eya proteins (Eya1, Eya2 and Eya3) to determine relationships between members of these 2 conserved families.

To understand further the properties of both the Six domain and the homeodomain it will be important to determine their 3 dimensional structures, particularly when bound to DNA. This has been done for several other homeodomain proteins by NMR spectroscopy, for example Antennapedia and MATα2 (Otting *et al.*, 1990; Wolberger *et al.*, 1991) and will help to answer the question of whether the Six domain is a DNA or protein binding domain or both. It is also important that the transcription activation domain of DMAHP is identified as has been done for AREC3/Six4 (Kawakami *et al.*, 1996a) as this will confirm its role as a transcription factor.

7.5 Genome Wide Screening for DNA Binding Targets of DMAHP

The whole genome PCR based screen identified 5 sequences to which DMAHP showed specific binding. From these, one candidate gene (dopamine D5 receptor gene) was identified by DNA sequencing and comparison to known genes in the databases. The other 4 sequences will require further characterisation in order to determine any genes with which they are associated. Although DMAHP was shown to bind to a transcribed region of the dopamine D5 receptor gene, transcription factors often bind upstream of the transcription start site and these sequences will not be present in cDNA libraries. Therefore, the next stage of this project will be to screen genomic DNA libraries using the selected sequences as probes, to obtain more genomic sequence. From this, known genes may be identified by database searching or novel genes can be characterised using gene structure prediction programs and by cDNA library screening using predicted coding regions as probes.

An alternative strategy to obtain more genomic sequence is *Alu*-exon PCR. This method involves performing PCR using a primer designed to bind to the end of an Alu repeat and a second primer that binds within the known genomic DNA (Buxton *et al.*, 1993). As *Alu* repeats occur, on average, every 4 kb in the human genome, there is a high probability of a product containing DNA from the gene of interest being amplified.

To demonstrate that the sequences selected can regulate the transcription of genes in cell culture they could be cloned into a vector upstream of a promoter that drives the expression of a chloramphenicol acetyltransferase (CAT) reporter gene. GST-Six+HD could then be subcloned into an expression vector that would allow the expression of the protein in the cell culture and the level of CAT activity, from the reporter gene,

measured. The chicken CDXA protein was shown to mediate an enhancing activity, measured by such a CAT assay, using sequences selected by whole genome PCR (Margalit *et al.*, 1993).

7.5.1 Alternative Strategies to Identify Binding Sites of DMAHP

Experiments described in chapter 5 of this thesis demonstrated the importance of a cofactor in enhancing the affinity of DMAHP for the ARE site. Therefore, the whole genome PCR screen may not have identified sequences to which DMAHP only binds in the presence of a cofactor or to which it only binds very weakly in the absence of the cofactor. An alternative strategy known as CASTing may be used to identify these sequences (Wright and Funk, 1993). CASTing uses crude nuclear extracts in place of purified protein, during the selection procedure, which leads to the selection of DNA sequences that preferentially bind multicomponent complexes. To identify sequences specific to a particular protein, antiserum raised against that protein is used to select for complexes containing it. This system was tested using a myogenin antibody to detect complexes formed after mixing crude myotube nuclear extracts with a random oligomer containing a degenerate core of 35 nucleotides. The clones were shown to contain the known myogenin heterodimer binding sites. Gel retardation assays using myotube nuclear extracts and selected sequences detected multiple bands. When analysed, the sequences were shown to contain consensus sequences, to which 6 factors with which myogenin interacts, are known to bind. This protocol could be adapted for use with nuclear cell extracts known to contain DMAHP. However, the only DMAHP antibodies available at present were raised against recombinant DMAHP fusion proteins and do not detect endogenously expressed human or mouse DMAHP protein (Charles Thornton personal communication and Winchester, 1997).

Although the majority of transcription factors bind untranscribed regions of the genome and can bind more than 500 bp upstream of the transcription start site of the gene which they are regulating, a method has been described for selecting for transcription factor binding sites close to or overlapping with transcribed DNA (Caubin *et al.*, 1994). The DNA selected by the whole genome PCR was hybridised with cDNA to select for sequences that were transcribed. Genes controlled by the thyroid hormone receptor were successfully selected using this system. However, a high percentage of down-regulated genes were identified despite reports suggesting that the majority of genes controlled by the thyroid hormone receptor are up-regulated. This is likely to be due to the fact that genes whose elements were greater than 500 bp upstream of their transcription start sites would not be detected by this method.

Therefore, this method is not appropriate for the identification of the majority of binding sites.

A system for the construction of a library of DNA-protein binding sites that contains sequences recognised by many of the DNA binding proteins present in a particular cell type has been reported (Nallur *et al.*, 1996). The library was constructed by the repeated selection of a random pool of oligonucleotides by gel retardation analysis with nuclear extracts. DNA present in retarded complexes was excised from the gel and used in a PCR reaction, the product of which was used in a second gel retardation assay. After 4 rounds of selection the sequences obtained were utilised as bait for the isolation of optimal sites for cloning transcription factors and therefore this system can investigate multiple transcription factors in a given cell type. Transcription factors involved upstream or downstream of DMAHP in the same pathway may be identified in this manner. This system, like that of CASTing has the advantage of using native protein and thus allows binding specificity to be modified by the presence of cofactors.

GST-Six+HD bound to the dopamine D5 receptor gene between the transcription and translation start sites which suggests that DMAHP could act as an RNA binding protein and therefore the RNA-protein binding properties of this protein should be investigated. It is possible to carry out whole genome PCR-based screens using RNA. The optimum RNA binding sites for T4 DNA polymerase were identified by using *in vitro* transcription to generate a library of RNA transcripts from double stranded oligonucleotides. Selected RNA was converted to cDNA for the PCR amplification step (Tuerk and Gold, 1990). Multiple approaches to the identification of target binding sites are advisable to enhance the likelihood of isolating all genes controlled by DMAHP.

7.6 Current Theories on the Mechanisms Involved in the Pathology of DM

Despite the identification of the mutation responsible for DM and the partial characterisation of 3 genes (*DMPK*, *DMAHP* and *59*) at the locus, the mechanism which causes the DNA expansion and the effect of the expansion on DNA/RNA/protein is still poorly understood.

7.6.1 Triplet Repeat Instability

To investigate the somatic and germline instability of the $(CTG)_n$ repeat, several groups have created transgenic mice carrying an expanded trinucleotide repeat integrated in a random location in the genome. It has been demonstrated that these repeats show intergenerational and somatic repeat instability and no correlation has been found between the somatic mutation rate and the proliferation rate of the tissues involved (Gourdon *et al.*, 1997; Lia *et al.*, 1998; Monckton *et al.*, 1997). This final result was surprising and suggests that if instability occurs during cell division some tissues have more efficient DNA repair systems than others. Alternatively instability may occur independently of cell division. These models require further analysis and should aid our understanding of the mechanisms involved in repeat instability.

The instability of the expanded $(CTG)_n$ repeat in cultured lymphoblastoid cell lines from DM patients was investigated by small pool PCR (Ashizawa *et al.*, 1996). Frequent mutations leading to small changes of repeat size were observed and it was suggested that they may represent the mechanism responsible for the somatic heterogeneity in the blood cells of DM patients. Rare mutations were shown to occur that lead to large changes of the $(CTG)_n$ repeat size, with a bias towards contractions.

Experiments using CAG/CTG-repeat tracts in yeast and *E.coli* have shown that the stability of these tracts is dependent on their orientation with respect to replication origins. The repeat was more unstable when the CTG tract was on the lagging strand (Freudenreich *et al.*, 1997; Wells, 1996). These data suggest a polymerase slippage model for the expansion of trinucleotide repeats. In this model dissociation and misaligned reannealing of the newly synthesised strand in the repetitive tract results in addition (or deletion) of repeats. CTG, CAG and CGG repeats are able to form hairpin structures and the lagging strand is single stranded for a longer period of time giving it more opportunity to form a hairpin and thus be mutated.

7.6.2 Effects of the Expansion

A $(CUG)_n$ triplet repeat pre-mRNA/mRNA binding protein has been isolated and characterised (Roberts *et al.*, 1997; Timchenko *et al.*, 1996). It is hypothesised that this protein (CUG-BP/hNab50) binds mRNA, regulating transcription/translation of a variety of genes. The increase in potential binding sites caused by the repeat expansion at the DM locus may reduce the amount of CUG-BP/hNab50 available to bind to other genes causing an RNA processing defect leading to some of the DM symptoms. It would also explain the accumulation of DM allele transcripts observed in the nucleus (Davis *et al.*, 1997). Recently, CUG-BP/hNab50 has been shown to bind to the

human cardiac troponin T (cTNT) pre-mRNA and regulate its alternative splicing. It was demonstrated that this splicing was disrupted in DM striated muscle and in normal cells expressing CUG repeats (Philips *et al.*, 1998). Therefore, altered expression of human cardiac troponin T may contribute to the cardiac defects seen in DM. A significant decrease in the insulin receptor RNA in both total RNA and RNA poly A^+ pools relative to normal and myopathic control muscles, measured relative to dystrophin and muscle sodium channel RNAs has been reported (Morrone *et al.*, 1997). As DM patients are known to show increased insulin resistance (Hudson *et al.*, 1987), it would be interesting to determine if CUG-BP/hNab50 binds to this RNA. Alternatively, the insulin receptor gene may be a target of DMAHP, which would also explain its downregulation in DM patients.

Two recent reports indicate that overexpression of *DMPK* inhibits the terminal differentiation of myoblasts (Okoli *et al.*, 1998; Sabourin *et al.*, 1997) and another contradictory report indicates that it increases a skeletal muscle phenotype in a muscle cell line (Bush *et al.*, 1996). However, the region responsible for the lack of myoblast differentiation was shown to be in the 3' UTR of *DMPK* upstream of the $(CTG)_n$ repeat (Sabourin *et al.*, 1997) which was not present in the construct which enhanced a skeletal muscle phenotype. These data further support the idea that an accumulation of *DMPK* mRNAs in the nucleus could be responsible for many of the symptoms associated with DM, in this case the delay of muscle maturation in congenital DM patients. However, it should also be remembered that the 3' UTR of *DMPK* is also the promoter region of *DMAHP*. Therefore, the inhibition of terminal differentiation may be due to the sequestering of proteins that normally regulate *DMAHP* and other genes in the muscle differentiation pathway.

The expanded $(CTG)_n$ repeat has been shown to alter chromatin structure thus leading to the downregulation of *DMAHP* (section 7.2) and hypermethylation has been shown to reduce the binding of the transcription factor Sp1 at a site proximal to expanded $(CTG)_n$ repeats (section 1.5.1). Therefore, alterations in chromatin structure have been demonstrated in regions that are controlling the expression of *DMAHP* and *DMPK*. In addition to *DMAHP*, *DMPK* and *59*, 3 further genes have been identified within 200 kb of the DM expanded repeat; gastric inhibitory polypeptide receptor gene (GIPR), 20D7, and SYMPLEKIN (Alwazzan *et al.*, 1998). It is important that the expression patterns and levels of these 3 genes are investigated in DM patients and normal controls to determine their role if any in the pathology of DM. 20D7 is of particular interest as it is known to be expressed in the testis and therefore may be involved in testicular atrophy and male infertility. GIPR is involved in the stimulation and regulation of insulin secretion in a glucose-dependent manner and might therefore be implicated in DM-associated diabetes.

7.7 Conclusion

Since the identification of the genetic defect responsible for DM many theories have been proposed and tested for the mechanism by which an expanded $(CTG)_n$ repeat in the 3'-UTR of one gene (*DMPK*) and the promoter region of a second gene (*DMAHP*) can lead to the diverse pathology associated with DM. It is becoming clear that the mechanism does not affect a single gene, but that the expanded $(CTG)_n$ repeat alters the local chromatin structure, which in turn affects the expression of several genes including *DMAHP*. It is possible that the greater the expansion, the more genes that are influenced. This could explain the increase in the severity of symptoms in successive generations, as the repeat expands. It is also likely that the expanded $(CUG)_n$ tract is sequestering RNA binding proteins, thus preventing them from binding to and correctly controlling the translation of many other genes which explains the multisystemic pathology of DM. The evidence for the involvement of *DMAHP* in the pathology of DM is now very strong. The work presented in this thesis has initiated the identification of genes involved in the pathways downstream of this transcription factor which are in turn likely to be involved in the pathogenesis of DM.

References

Abdelhak, S., Kalatzis, V., Heilig, R., Compain, S., Samson, D., Vincent, C., Weil, D., Cruaud, C., Sahly, I., Leibovici, M., Bitner Glindzicz, M., Francis, M., Lacombe, D., Vigneron, J., Charachon, R., Boven, K., Bedbeder, P., Van Regemorter, N., Weissenbach, J. and Petit, C. (1997). A human homologue of the Drosophila eyes absent gene underlies Branchio Oto Renal (BOR) syndrome and identifies a novel gene family. Nat. Genet. **15**, 157-164.

Affolter, M., Percival-Smith, A., Muller, M., Leupin, W. and Gehring, W. J. (1990). DNA binding properties of the purified Antennapedia homeodomain. Proc. Natl. Acad. Sci. (USA) **87**, 4093-4097.

Akarsu, A. N., Stolivov, I., Yilmaz, E., Sayli, B. S. and Sarfarazi, M. (1996). Genomic structure of *HOXD13* gene: a nine polyalanine duplication causes synpolydactyly in two unrelated families. Hum. Mol. Genet. **5**, 945-952.

Alazard, R., Betermier, M. and Chandler, M. (1992). *Escherichia coli* integration host factor stabilises bacteriophage Mu repressor interactions with operator DNA in vitro. Molecular Microbiology **6**, 1707-1714.

Alwazzan, M., Hamshere, M. G., Lennon, G. G. and Brook, J. D. (1998). Six transcripts map within 200 kilobases of the myotonic dystrophy expanded repeat. Mammalian Genome **9**, 485-487.

Andersen, B. and Rosenfeld, M. G. (1994). Pit-1 determines cell types during development of the anterior pituitary gland. J. Biol. Chem. **269**, 29335-29338.

Anvret, M., Ahlberg, G., Grandell, U., Hedberg, B., Johnson, K. and Edstrom, L. (1993). Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. Hum. Mol. Genet. **2**, 1397-1400.

Ashizawa, T., Monckton, D. G., Vaishnav, S., Patel, B. J., Voskova, A. and Caskey, C. T. (1996). Instability of the expanded $(CTG)_n$ repeats in the myotonin protein kinase gene in cultured lymphoblastoid cell lines from patients with myotonic dystrophy. Genomics **36**, 47-53.

Aslanidis, C., Jansen, G., Amemiya, C., Shutler, G., Mahadevan, M., Tsilfidis, C., Chen, C., Alleman, J., Wormskamp, N. G. M., Vooijs, M., Buxton, J., Johnson, K., Smeets, H. J. M., Lennon, G. G., Carrano, A. V., Korneluk, R. G., Wieringa, B. and de Jong, P. J. (1992). Cloning of the essential myotonic dystrophy region and mapping of the putative defect. Nature **355**, 548-551.

Baldwin, C. T., Hoth, C. F., Amos, J. A., Dasilva, E. O. and Milunsky, A. (1992). An exonic mutation in the *HuP2* paired domain gene causes Waardenburg's syndrome. Nature **355**, 637-638.

Batten, F. E. and Gibb, H. P. (1909). Myotonia atrophica. Brain **32**, 187-205.

Bednarz, A. L., Boocock, M. R. and Sherratt, D. J. (1990). Determinants of correct *res* site alignment in site-specific recombination by Tn3 resolvase. Genes and Development **4**, 2366-2375.

Behrens, M. I., Jalil, P., Serani, A., Vergara, F. and Alvarez, O. (1994). Possible role of apamin-sensitive K^+ channels in myotonic dystrophy. Muscle & Nerve **17**, 1264-1270.

Beischlag, T. V., Marchese, A., Meador-Woodruff, J. H., Damask, S. P., O'Dowd, B. F., Tyndale, R. F., Van Tol, H. H. M., Seeman, P. and Niznik, H. B. (1995). The human dopamine D5 receptor gene: Cloning and characterization of the 5'-flanking and promoter region. Biochemistry **34**, 5960-5970.

Benders, A., Wevers, R. A. and Veerkamp, J. H. (1996). Ion transport in human skeletal muscle cells disturbances in myotonic dystrophy and Brody's disease. Acta Physiol. Scand. **156**, 355-367.

Benders, A. A. G. M., Timmermans, J. A. H., Oosterhof, A., Ter Laak, H. J., van Kuppevelt, T. H. M. S. M., Wevers, R. A. and Veerkamp, J. H. (1993). Deficiency of Na^+/K^+-ATPase and sarcoplasmic reticulum Ca^{2+}-ATPase in skeletal muscle and cultured muscle cells of myotonic dystrophy patients. Biochem. J. **293**, 269-274.

Billeter, M., Qian, Y. Q., Otting, G., Muller, M., Gehring, W. and Wuthrich, K. (1993). Determination of the nuclear magnetic resonance solution structure of an *Antennapedia* Homeodomain-DNA complex. J. Mol. Biol. **234**, 1084-1093.

Birnboim, H. C. and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Research **7**, 1513-1523.

Boncinelli, E. (1997). Homeobox genes and disease. Curr. Opin. Genet. Dev. **7**, 331-337.

Bopp, D., Burri, M., Baumgartner, G., Frigerio, G. and Noll, M. (1986). Conservation of a large protein domain in the segmentation gene *paired* and in functionally related genes in *Drosophila*. Cell **47**, 1033-1049.

Boucher, C. A. (1996). Identification and characterisation of two human homeodomain encoding genes: *DMAHP* and *SIX1*. Unpublished PhD, London University, London.

Boucher, C. A., Carey, N., Edwards, Y. H., Siciliano, M. J. and Johnson, K. J. (1996). Cloning of the human *SIX1* gene and its assignment to chromosome 14. Genomics **33**, 140-142.

Boucher, C. A., King, S. K., Carey, N., Krahe, R., Winchester, C. L., Rahman, S., Creavin, T., Meghji, P., Bailey, M. E. S., Chartier, F. L., Brown, S. D., Siciliano, M. J. and Johnson, K. J. (1995). A novel homeodomain encoding gene is associated with a large CpG island interrupted by the myotonic dystrophy unstable $(CTG)_{(n)}$ repeat. Hum. Mol. Genet. **4**, 1919-1925.

Bovolenta, P., Mallamaci, A. and Boncinelli, E. (1996). Cloning and characterisation of two chick homeobox genes, members of the Six/sine oculis family, expressed during eye development. Int. J. Dev. Biol. **1**, 73-74.

Bovolenta, P., Mallamaci, A., Puelles, L. and Boncinelli, E. (1998). Expression pattern of *cSix3*, a member of the Six/sine oculis family of transcription factors. Mech. Dev. **70**, 201-203.

Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J. P., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P. S., Shaw, D. J. and Housman, D. E. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell **68**, 799-808.

Brunner, H. G., Nillesen, W., van Oost, B. A., Jansen, G., Wieringa, B., Ropers, H. H. and Smeets, H. J. M. (1992). Presymptomatic diagnosis of myotonic dystrophy. J. Med. Genet. **29**, 780-784.

Bush, E. W., Taft, C. S., Meixell, G. E. and Perryman, M. B. (1996). Overexpression of myotonic dystrophy kinase in BC(3)H1 cells induces the skeletal muscle phenotype. J. Biol. Chem. **271**, 548-552.

Buxton, J., Davies, J., Shelbourne, P., Yokobata, K., Williamson, R. and Johnson, K. (1993). Isolation and ordering of bacteriophage genomic clones corresponding to 2 yacs from 19q13.3. Molecular and Cellular Probes **7**, 75-80.

Buxton, J., Shelbourne, P., Davies, J., Jones, C., Van Tongeren, T., Aslanidis, C., de Jong, P., Jansen, G., Anvret, M., Riley, B., Williamson, R. and Johnson, K. (1992). Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. Nature **355**, 547-548.

Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Canizares, J., Koutnikova, H., Bidichandani, S. I., Gellera, C., Brice, A., Trouillas, P., Demichele, G., Filla, A., Defrutos, R., Palau, F., Patel, P. I., Didonato, S., Mandel, J. L., Cocozza, S., Koenig, M. and Pandolfo, M. (1996). Friedreich's Ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science **271**, 1423-1427.

Carango, P., Noble, J. E., Marks, H. G. and Funanage, V. L. (1993). Absence of myotonic dystrophy protein kinase (DMPK) mRNA as a result of a triplet repeat expansion in myotonic dystrophy. Genomics **18**, 340-348.

Carey, N., Johnson, K., Nokelainen, P., Peltonen, L., Savontaus, M. L., Juvonen, V., Anvret, M., Grandell, U., Chotai, K., Robertson, E., Middletonprice, H. and Malcolm, S. (1994). Meiotic drive at the myotonic dystrophy locus. Nat. Genet. **6**, 117-118.

Caubin, J., Iglesias, T., Bernal, J., Munoz, A., Marquez, G., Barbero, J. L. and Zaballos, A. (1994). Isolation of genomic DNA fragments corresponding to genes modulated in vivo by a transcription factor. Nucleic Acids Research **22**, 4132-4138.

Chakraborty, R., Stivers, D. N., Deka, R., Yu, L. M., Shriver, M. D. and Ferrell, R. E. (1996). Segregation distortion of the CTG repeats at the myotonic dystrophy locus. Am. J. Hum. Genet. **59**, 109-118.

Cheyette, B. N. R., Green, P. J., Martin, K., Garren, H., Hartenstein, V. and Zipursky, S. L. (1994). The Drosophila *sine oculis* locus encodes a homeodomain-containing protein required for the development of the entire visual system. Neuron **12**, 977-996.

Chodosh, L. A., Carthew, R. W. and Sharp, P. A. (1986). A single polypeptide possesses the binding and transcription activities of the adenovirus major late transcription factor. Mol. Cell. Biol. **6**, 4723-4733.

David, G., Abbas, N., Stevanin, G., Durr, A., Yvert, G., Cancel, D., Weber, C., Imbert, G., Saudou, F., Antoniou, E., Drabkin, H., Gemmill, R., Giunti, P., Benomar, A., Wood, N., Ruberg, M., Agid, Y., Mandel, J. L. and Brice, A. (1997). Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. Nat. Genet. **17**, 65-70.

Davies, J., Yamagata, H., Shelbourne, P., Buxton, J., Ogihara, T., Nokelainen, P., Nakagawa, M., Williamson, R., Johnson, K. and Miki, T. (1992). Comparison of the myotonic dystrophy associated CTG repeat in European and Japanese populations. J. Med. Genet. **29**, 766-769.

Davis, B. M., McCurrach, M. E., Taneja, K. L., Singer, R. H. and Housman, D. E. (1997). Expansion of a CUG trinucleotide repeat in the 3' untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts. Proc. Natl. Acad. Sci. (USA) **94**, 7388-7393.

De Kok, Y. J. M., Merkx, G. F. M., Vav der Maarel, S. M., Huber, I., Malcolm, S., Ropers, H. H. and Cremers, F. P. M. (1995). A duplication /paracentric inversion associated with familial X-linked deafness (DFN3) suggests the presence of a regulatory element more than 400 kb upstream of the *POU3F4* gene. Hum. Mol. Genet. **4**,

Denhardt, D. T. (1966). A membrane-filter technique for the detection of complementary DNA. Biochem. Biophys. Res. Commun. **23**, 641-646.

Desnuelle, C., Lombet, A., Serratrice, G. and Lazdunski, M. (1982). Sodium-channel and sodium-pump in normal and pathological muscles from patients with myotonic muscular dystrophy and lower motor neuron impairment. J. Clin. Invest. **69**, 358-367.

Driever, W. and Nusslein-Volhard, C. (1989). The bicoid protein is a positive regulator of *hunchback* transcription in the early *Drosophila* embryo. Nature **337**, 138-143.

Dubnau, J. and Struhl, G. (1996). RNA recognition and translational regulation by a homeodomain protein. Nature **379**, 694-699.

Dunne, P. W., Ma, L., Casey, D. L., Harati, Y. and Epstein, H. F. (1996). Localization of myotonic dystrophy protein kinase in skeletal muscle and its alteration with disease. Cell Motil. Cytoskeleton **33**, 52-63.

Edstrom, L. and Wroblewski, R. (1989). Intracellular elemental composition of single muscle fibers in muscular dystrophy and dystrophia myotonica. Acta Neurol. Scand. **80**, 419-424.

Ekker, S. C., Young, K. E., von Kessler, D. P. and Beachy, P. A. (1991). Optimal DNA sequence recognition by the Ultrabithorax homeodomain of *Drosophila*. Embo J. **10**, 1179-1186.

Engelkamp, D. and van Heyningen, V. (1996). Transcription factors in disease. Curr. Opin. Genet. Dev. **6**, 334-342.

Etongue-Mayer, P., Faure, R., Bouchard, J. P. and Puymirat, J. (1998). Characterization of a 54 kilodalton human protein kinase recognized by an antiserum raised against the myotonin kinase. Muscle & Nerve **21**, 8-17.

Fainsod, A., Margalit, Y., Haffner, R. and Gruenbaum, Y. (1991). Non-immunological precipitation of protein-DNA complexes using glutathione-S-transferase fusion proteins. Nucleic Acids Research **19**, 4005.

Fields, S. and Sternglanz, R. (1994). The two-hybrid system: an assay for protein-protein interactions. Trends in Genetics **10**, 286-292.

Freudenreich, C. H., Stavenhagen, J. B. and Zakian, V. A. (1997). Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Mol. Cell. Biol. **17**, 2090-2098.

Fu, Y. H., Friedman, D. L., Richards, S., Pearlman, J. A., Gibbs, R. A., Pizzuti, A., Ashizawa, T., Perryman, M. B., Scarlato, G., Fenwick, R. G. and Caskey, C. T. (1993). Decreased expression of myotonin protein-kinase messenger RNA and protein in adult form of myotonic dystrophy. Science **260**, 235-238.

Fu, Y. H., Pizzuti, A., Fenwick, R. G., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., de Jong, P., Wieringa, B., Korneluk, R., Perryman, M. B., Epstein, H. F. and Caskey, C. T. (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science **255**, 1256-1258.

Gehring, W. J., Affolter, M. and Burglin, T. (1994). Homeodomain proteins. Ann. Rev. Biochem. **63**, 487-526.

Gennarelli, M., Dallapiccola, B., Baiget, M., Martorell, L. and Novelli, G. (1994). Meiotic drive at the myotonic dystrophy locus. J. Med. Genet. **31**, 980-980.

Goldman, A., Ramsay, M. and Jenkins, T. (1995). New founder haplotypes at the myotonic dystrophy locus in southern Africa. Am. J. Hum. Genet. **56**, 1373-1378.

Gourdon, G., Radvanyi, F., Lia, A. S., Duros, C., Blanche, M., Abitbol, M., Junien, C. and HofmannRadvanyi, H. (1997). Moderate intergenerational and somatic instability of a 55 CTG repeat in transgenic mice. Nat. Genet. **15**, 190-192.

Goutte, C. and Johnson, A. D. (1988). a1 protein alters the DNA binding specificity of α2 repressor. Cell **52**,

Goutte, C. and Johnson, A. D. (1993). Yeast a1 and α2 homeodomain proteins form a DNA-binding activity with properties distinct from those of either protein. J. Mol. Biol. **233**, 359-371.

Gruss, P. and Walther, C. (1992). Pax in development. Cell **69**, 719-722.

Guan, K. L. and Dixon, J. E. (1991). Eukaryotic proteins expressed in *Escherichia Coli*: an improved thrombin cleavage and purification procedure of fusion proteins with glutathione S transferase. Analyt. Biochem. **192**, 262-267.

Hamshere, M. G., Newman, E. E., Alwazzan, M., Athwal, B. S. and Brook, J. D. (1997). Transcriptional abnormality in myotonic dystrophy affects *DMPK* but not neighboring genes. Proc. Natl. Acad. Sci. (USA) **94**, 7394-7399.

Hanes, S. D. and Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix 9. Cell **57**, 1275-1283.

Hanson, I. and van Heyningen, V. (1995). Pax6: More than meets the eye. Trends in Genetics **11**, 268-272.

Hanson, I. M., Seawright, A., Hardman, K., Hodgson, S., Zaletayev, D., Fekete, G. and van Heyningen, V. (1993). Pax6 mutations in aniridia. Hum. Mol. Genet. **2**, 915-920.

Harley, H. G., Rundle, S. A., Reardon, W., Myring, J., Crow, S., Brook, J. D., Harper, P. S. and Shaw, D. J. (1992). Unstable DNA-sequence in myotonic dystrophy. Lancet **339**, 1125-1128.

Harper, P. S. (1989). Myotonic Dystrophy (2nd ed.) W B Saunders Company.

Harper, P. S., Harley, H. G., Reardon, W. and Shaw, D. J. (1992). Anticipation in myotonic dystrophy - New light on an old problem. Am. J. Hum. Genet. **51**, 10-16.

Hazbun, T. R., Stahura, F. L. and Mossing, M. C. (1997). Site-specific recognition by an isolated DNA-binding domain of the sine oculis protein. Biochemistry **36**, 3680-3686.

Heath, S. K., Carne, S., Hoyle, C., Johnson, K. J. and Wells, D. J. (1997). Characterisation of expression of *mDMAHP* a homeodomain encoding gene at the murine DM locus. Hum. Mol. Genet. **6**, 651-657.

Herr, W., Sturm, R. A., Clerc, R. G., Corcoran, L. M., Baltimore, D., Sharp, P. A., Ingraham, H. A., Rosenfeld, M. G., Finney, M., Ruvkun, G. and Horvitz, H. R. (1988). The POU domain: a large conserved region in the mammalian *pit*-1, *oct*-1, *oct*-2 and *Caenorhabditis elegans unc*-86 gene products. Genes and Development **2**, 1513-1516.

Hertz, G. Z., Hartzell III, G. W. and Stormo, G. D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. CABIOS **6**, 81-92.

Hoey, T. and Levine, M. (1988). Divergent homeo box proteins recognize similar DNA sequences in *Drosophila*. Nature **332**, 858-861.

Hofmann-Radvanyi, H., Lavedan, C., Rabes, J.-P., Savoy, D., Duros, C., Johnson, K. and Junien, C. (1993). Myotonic dystrophy: absence of CTG enlarged

transcript in congenital forms, and low expression of the normal allele. Hum. Mol. Genet. **2**, 1263-1266.

Hudson, A. J., Huff, M. W., Wright, C. G., Silver, M. M., Lo, T. C. Y. and Banerjee, D. (1987). The role of insulin resistance in the pathogenesis of myotonic muscular dystrophy. Brain **110**, 469-488.

Hurst, G. D. D., Hurst, L. D. and Barrett, J. A. (1995). Meiotic drive and myotonic dystrophy. Nat. Genet. **10**, 132-133.

Imbert, G., Kretz, C., Johnson, K. and Mandel, J.-L. (1993). Origin of the expansion mutation in myotonic dystrophy. Nat. Genet. **4**, 72-76.

Imbert, G., Saudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J. M., Weber, C., Mandel, J. L., Cancel, G., Abbas, N., Durr, A., Didierjean, O., Stevanin, G., Agid, Y. and Brice, A. (1996). Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. Nat. Genet. **14**, 285-291.

Ingraham, H. A., Flynn, S. E., Voss, J. W., Albert, V. R., Kapiloff, M. S., Wilson, L. and Rosenfeld, M. G. (1990). The POU-specific domain of Pit-1 is essential for sequence-specific, high affinity DNA binding and DNA-dependent Pit-1-Pit-1 interactions. Cell **61**, 1021-1033.

Inoue, S., Orimo, A., Hosoi, T., Kondo, S., Toyoshima, H., Kondo, T., Ikegami, A., Ouchi, Y., Orimo, H. and Murasmatsu, M. (1993). Genomic binding-site cloning reveals an estrogen-responsive gene that encodes a RING finger protein. Proc. Natl. Acad. Sci. (USA) **90**, 11117-11121.

Ish-Horowicz, D. and Burke, J. F. (1981). Rapid and efficient cosmid cloning. Nucleic Acids Research **9**, 2989.

Israel, D. I. (1993). A PCR-based method for high stringency screening of DNA libraries. Nucleic Acids Research **21**, 2627-2631.

Jansen, G., Bachner, D., Coerwinkel, M., Wormskamp, N., Hameister, H. and Wieringa, B. (1995). Structural organization and developmental expression pattern of the mouse WD-repeat gene DMR-N9 immediately upstream of the myotonic dystrophy locus. Hum. Mol. Genet. **4**, 843-852.

Jansen, G., Groenen, P., Bachner, D., Jap, P. H. K., Coerwinkel, M., Oerlemans, F., Vandenbroek, W., Gohlsch, B., Pette, D., Plomp, J. J., Molenaar, P. C., Nederhoff, M. G. J., Vanechteld, C. J. A., Dekker, M., Berns, A., Hameister, H. and Wieringa, B. (1996). Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice. Nat. Genet. **13**, 316-324.

Jansen, G., Mahadevan, M., Amemiya, C., Wormskamp, N., Segers, B., Hendriks, W., O'hoy, K., Baird, S., Sabourin, L., Lennon, G., Jap, P. L., Iles, D., Coerwinkel, M., Hofker, M., Carrano, A. V., de Jong, P. J., Korneluk, R. G. and Wieringa, B. (1992). Characterization of the myotonic dystrophy region predicts multiple protein isoform-encoding mRNAs. Nat. Genet. **1**, 261-266.

Jansen, G., Willems, P., Coerwinkel, M., Nillesen, W., Smeets, H., Vits, L., Howeler, C., Brunner, H. and Wieringa, B. (1994). Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic events in (CTG)n repeat variation and selection against extreme expansion in sperm. Am. J. Hum. Genet. **54**, 575-585.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S. and Kakizuka, A. (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat. Genet. **8**, 221-228.

Kawakami, K., Ohto, H., Ikeda, K. and Roeder, R. G. (1996a). Structure, function and expression of a murine homeobox protein AREC3, a homologue of *Drosophila sine oculis* gene product, and implication in development. Nucleic Acids Research **24**, 303-310.

Kawakami, K., Ohto, H., Takizawa, T. and Saito, T. (1996b). Identification and expression of *six* family genes in mouse retina. Febs Letters **393**, 259-263.

Kelly, M., Burke, J., Smith, M., Klar, A. and Beach, D. (1988). Four mating-type genes control sexual differentiation in the fission yeast. Embo J. **7**, 1537-1547.

Kessel, M. and Gruss, P. (1990). Murine developmental control genes. Science **249**, 374-379.

King, S. K. (1996). Gene expression across the murine myotonic dystrophy (DM) region. Unpublished PhD, London University, London.

Kinzler, K. W. and Voglestein, B. (1989). Whole genome PCR: application to the identification of sequences bound by gene reguatory proteins. Nucleic Acids Research **17**, 3645-3653.

Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B. and Pabo, C. O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: A framework for understanding homeodomain-DNA interactions. Cell **63**, 579-590.

Kleinschmidt, C., Tovar, K. and Hillen, W. (1991). Computer simulations and experimental studies of gel mobility patterns for weak and strong non-cooperative protein binding to two targets on the same DNA: application of binding to Tet

repressor variants to multiple and single *tet* operator sites. Nucleic Acids Research **19**, 1021-1028.

Klesert, T. R., Otten, A. D., Bird, T. D. and Tapscott, S. J. (1997). Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of *DMAHP*. Nat. Genet. **16**, 402-406.

Knight, S. J. L., Flannery, A. V., Hirst, M. C., Campbell, L., Christodoulou, Z., Phelps, S. R., Pointon, J., Middleton-Price, H. R., Barnicoat, A., Pembrey, M. E., Holland, J., Oostra, B. A., Bobrow, M. and Davies, K. E. (1993). Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. Cell **74**, 127-134.

Kobayashi, M., Toyama, R., Takeda, H., Dawid, I.B., Kawakami, K. (1998). Overexpression of the forebrain-specific homeobox gene *six3* induces rostral forebrain enlargment in zebrafish. Development **125**, 2973-2982.

Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F. and Tsuji, S. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral pallidoluysian atrophy (DRPLA). Nat. Genet. **6**, 9-13.

Krahe, R., Ashizawa, T., Abbruzzese, C., Roeder, E., Carango, P., Giacanelli, M., Funanage, V. L. and Siciliano, M. J. (1995a). Effect of myotonic dystrophy trinucleotide repeat expansion on DMPK transcription and processing. Genomics **28**, 1-14.

Krahe, R., Eckhart, M., Ogunniyi, A. O., Osuntokun, B. O., Siciliano, M. J. and Ashizawa, T. (1995b). De novo myotonic dystrophy mutation in a Nigerian kindred. Am. J. Hum. Genet. **56**, 1067-1074.

La Spada, R. A., Wilson, E. M., Lubahn, D. B., Harding, A. E. and Fischbeck, K. H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature **352**, 77-79.

Laemmli, E. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature **227**, 680-685.

Latchman, D. S. (1998). Eukaryotic transcription factors (3rd ed.) Academic Press.

Lavedan, C., Hofmann-Radvanyi, H., Shelbourne, P., Rabes, J.-P., Duros, C., Savoy, D., Dehaupas, I., Luce, S., Johnson, K. and Junien, C. (1993). Myotonic dystrophy: size- and sex-dependent dynamics of CTG meiotic instability, and somatic mosaicism. Am. J. Hum. Genet. **52**, 875-883.

Leeflang, E. P., McPeek, M. S. and Arnheim, N. (1996). Analysis of meiotic segregation, using single sperm typing: Meiotic drive at the myotonic dystrophy locus. Am. J. Hum. Genet. **59**, 896-904.

Lia, A. S., Seznec, H., Radvanyi, H., Radvanyi, F., Duros, C., Saquet, C., Blanche, M., Junien, C. and Gourdon, G. (1998). Somatic instability of the CTG repeat in mice transgenic for the myotonic dystrophy region is age dependent but not correlated to the relative intertissue transcription levels and proliferative capacities. Hum. Mol. Genet. **7**, 1285-1291.

Loosli, F., Koster, R. and Wittbrodt, J. (1997). The medaka homologue of the sine oculis gene is expressed in the developing eye. Developmental Biology **186**, B24.

Love, J. J., Li, X., Case, D. A., Giese, K., Grosschedl, R. and Wright, P. E. (1995). Structural basis for DNA bending by the architectural transcription factor LEF-1. Nature **376**, 791-795.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N., Macfarlane, H., Jenkins, B., Anderson, M. A., Wexler, N. S., Gusella, J. F., Bates, G. P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A. M., Lehrach, H., Buckler, A. J., Church, D., Doucettestamm, L., Odonovan, M. C., Ribaramirez, L., Shah, M., Stanton, V. P., Strobel, S. A., Draths, K. M., Wales, J. L., Dervan, P., Housman, D. E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J. J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F. S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D. and Harper, P. S. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell **72**, 971-983.

Maeda, M., Taft, C. S., Bush, E. W., Holder, E., Bailey, W. M., Neville, H., Perryman, M. B. and Bies, R. D. (1995). Identification, tissue-specific expression, and subcellular- localization of the 80-kDa and 71-kDa forms of myotonic dystrophy kinase protein. J. Biol. Chem. **270**, 20246-20249.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'hoy, K., Leblond, S., Earle-MacDonald, J., de Jong, P. J., Wieringa, B. and Korneluk, R. G. (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. Science **255**, 1253-1255.

Mahadevan, M. S., Amemiya, C., Jansen, G., Sabourin, L., Baird, S., Neville, C. E., Wormskamp, N., Segers, B., Batzer, M., Lamerdin, J., de Jong, P., Wieringa,

B. and Korneluk, R. G. (1993a). Structure and genomic sequence of the myotonic dystrophy (DM kinase) gene. Hum. Mol. Genet. **2**, 299-304.

Mahadevan, M. S., Foitzik, M. A., Surh, L. C. and Korneluk, R. G. (1993b). Characterization and polymerase chain reaction (PCR) detection of an *Alu* deletion polymorphism in total linkage disequilibrium with myotonic dystrophy. Genomics **15**, 446-448.

Mann, R. S. and Chan, S.-K. (1996). Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. Trends in Genetics **12**, 258-262.

Margalit, Y., Yarus, S., Shapira, E., Gruenbaum, Y. and Fainsod, A. (1993). Isolation and characterization of target sequences of the chicken *CdxA* homeobox gene. Nucleic Acids Research **21**, 4915-4922.

Martorell, L., Martinez, J. M., Carey, N., Johnson, K. and Baiget, M. (1995). Comparison of CTG repeat length expansion and clinical progression of myotonic dystrophy over a five year period. J. Med. Genet. **32**, 593-596.

Mead, J., Zhong, H., Acton, T. B. and Vershon, A. K. (1996). The yeast α2 and Mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes. Mol. Cell. Biol. **16**, 2135-2143.

Mirzayans, F., Pearce, W. G., Macdonald, I. M. and Walter, M. A. (1995). Mutation of the *PAX6* gene in patients with autosomal dominant keratitis. Am. J. Hum. Genet. **57**, 539-548.

Monckton, D. G., Coolbaugh, M. I., Ashizawa, K. T., Siciliano, M. J. and Caskey, C. T. (1997). Hypermutable myotonic dystrophy CTG repeats in transgenic mice. Nat. Genet. **15**, 193-196.

Monckton, D. G., Wong, L. J. C., Ashizawa, T. and Caskey, C. T. (1995). Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. Hum. Mol. Genet. **4**, 1-8.

Morinaga, T., Yasuda, H., Hashimoto, T., Higashio, K. and Tamaoki, T. (1991). A human α- fetoprotein enhancer-binding protein, ATBF1, contains four homeodomains and seventeen zinc fingers. Mol. Cell. Biol. **11**, 6041-6049.

Morrone, A., Pegoraro, E., Angelini, C., Zammarchi, E., Marconi, G. and Hoffman, E. P. (1997). RNA metabolism in myotonic dystrophy: Patient muscle shows decreased insulin receptor RNA and protein consistent with abnormal insulin resistance. J. Clin. Invest. **99**, 1691-1698.

Mortlock, D. P. and Innis, J. W. (1997). Mutation of *HOXA13* in hand-foot-genital syndrome. Nat. Genet. **15**, 179-181.

Muller, M., Affolter, M., Leupin, W., Otting, G., Wuthrich, K. and Gehring, W. J. (1988). Isolation and sequence-specific DNA binding of the *Antennapedia* homeodomain. Embo J. **7**, 4299-4304.

Murakami, Y., Ohto, H., Ikeda, U., Shimada, K., Momoi, T. and Kawakami, K. (1998). Promoter of *mDMAHP/Six5*: differential utilization of multiple transcription initiation sites and positive/negative regulatory elements. Hum. Mol. Genet. **7**, 2103-2112.

Nakagawa, M., Yamada, H., Higuchi, I., Kaminishi, Y., Miki, T., Johnson, K. and Osame, M. (1994). A case of paternally inherited congenital myotonic dystrophy. J. Med. Genet. **31**, 397-400.

Nallur, G. N., Prakash, K. and Weissman, S. M. (1996). Multiplex selection technique (MuST): An approach to clone transcription factor binding sites. Proc. Natl. Acad. Sci. (USA) **93**, 1184-1189.

Nguyen, T., Bard, J., Jin, H., Tarusico, D., Ward, D. C., Kennedy, J. L., Weinshank, R., Seeman, P. and O'Dowd, B. F. (1991). Human dopamine D5 receptor pseudogenes. Gene **109**, 211-218.

Neville, C. E., Mahadevan, M. S., Barcelo, J. M. and Korneluk, R. G. (1994). High resolution genetic analysis suggests one ancestral predisposing haplotype for the origin of the myotonic dystrophy mutation. Hum. Mol. Genet. **3**, 45-51.

Ohto, H., Takizawa, T., Saito, T., Kobayashi, M., Ikeda, K. and Kawakami, K. (1998). Tissue and developmental distribution of *Six* family gene products. Int. J. Dev. Biol. **42**, 141-148.

Okoli, G., Carey, N., Johnson, K. J. and Watt, D. J. (1998). Over expression of the murine myotonic dystrophy protein kinase in the mouse myogenic C2C12 cell line leads to inhibition of terminal differentiation. Biochem. Biophys. Res. Comm. **246**, 905-911.

Oliver, G. and Gruss, P. (1997). Current views on eye development. Trends in Neurosciences **20**, 415-421.

Oliver, G., Mailhos, A., Wehr, R., Copeland, N. G., Jenkins, N. A. and Gruss, P. (1995a). *Six3*, a murine homologue of the *sine oculis* gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development. Development **121**, 4045-4055.

Oliver, G., Wehr, R., Jenkins, N. A., Copeland, N. G., Cheyette, B. N. R., Hartenstein, V., Zipursky, S. L. and Gruss, P. (1995b). Homeobox genes and connective tissue patterning. Development **121**, 693-705.

Orr, H. T., Chung, M. Y., Banfi, S., Kwiatkowski, T. J., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P. W. and Zoghbi, H. Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type-1. Nat. Genet. **4**, 221-226.

Otten, A. D. and Tapscott, S. J. (1995). Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure. Proc. Natl. Acad. Sci. (USA) **92**, 5465-5469.

Otting, G., Qian, Y. Q., Billeter, M., Muller, M., Affolter, M., Gehring, W. J. and Wuthrich, K. (1990). Protein-DNA contacts in the structure of a homeodomain-DNA complex determined by nuclear magnetic resonance spectroscopy in solution. Embo J. **9**, 3085-3092.

Otting, G., Qian, Y. Q., Muller, M., Affolter, M., Gehring, W. and Wuthrich, K. (1988). Secondary structure determination for the *Antennapedia* homeodomain by nuclear magnetic resonance and evidence for a helix-turn-helix motif. Embo J. **7**, 4305-4309.

Pellegrino, M., Pellegrini, M., Bigini, P. and Scimemi, A. (1998). Properties of Ca^{2+}-activated K^{+} channels in erythrocytes from patients with myotonic muscular dystrophy. Muscle & Nerve **21**, 1465-1472.

Pham, Y. C. N., Man, N. t., Lam, L. T. and Morris, G. E. (1998). Localisation of myotonic dystrophy protein kinase in human and rabbit tissues using a new panel of monoclonal antibodies. Hum. Mol. Genet. **7**, 1957-1965.

Philips, A. V., Timchenko, L. T. and Cooper, T. A. (1998). Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. Science **280**, 737-741.

Pieretti, M., Zhang, F., Fu, Y.-H., Warren, S. T., Oostra, B. A., Caskey, C. T. and Nelson, D. L. (1991). Absence of expression of the *FMR-1* gene in fragile X syndrome. Cell **66**, 817-822.

Pignoni, F., Hu, B. R., Zavitz, K. H., Xiao, J. A., Garrity, P. A. and Zipursky, S. L. (1997). The eye specification proteins so and eya form a complex and regulate multiple steps in Drosophila eye development. Cell **91**, 881-891.

Qian, Y. Q., Billeter, M., Otting, G., Muller, M., Gehring, W. J. and Wuthrich, K. (1989). The structure of the *Antennapedia* homeodomain determined by NMR spectroscopy in solution: comparison with prokaryotic repressors. Cell **59**, 573-580.

Reddy, S., Smith, D. B. J., Rich, M. M., Leferovich, J. M., Reilly, P., Davis, B. M., Tran, K., Rayburn, H., Bronson, R., Cros, D., Balicegordon, R. J. and Housman, D. (1996). Mice lacking the myotonic dystrophy protein kinase develop a late onset progressive myopathy. Nat. Genet. **13**, 325-335.

Renaud, J. F., Desnuelle, C., Schmidantomarchi, H., Hugues, M., Serratrice, G. and Lazdunski, M. (1986). Expression of apamin receptor in muscles of patients with myotonic muscular dystrophy. Nature **319**, 678-680.

Rice, P. A., Yang, S.-w., Mizuuchi, K. and Nash, H. A. (1996). Crystal structure of an IHF-DNA complex: A protein-induced DNA U-turn. Cell **87**, 1295-1306.

Roberts, R., Timchenko, N. A., Miller, J. W., Reddy, S., Caskey, C. T., Swanson, M. S. and Timchenko, L. T. (1997). Altered phosphorylation and intracellular distribution of a $(CUG)_n$ triplet repeat RNA-binding protein in patients with myotonic dystrophy and in myotonin protein kinase knockout mice. Proc. Natl. Acad. Sci. (USA) **94**, 13221-13226.

Roses, A. D. and Appel, S. H. (1973). Protein kinase activity in erythrocyte ghosts of patients with myotonic muscular dystrophy. Proc. Natl. Acad. Sci. (USA) **70**, 1855-1859.

Roses, A. D. and Appel, S. H. (1974). Muscle membrane protein kinase in myotonic muscular dystrophy. Nature **250**, 245-247.

Rubinsztein, D. C., Leggo, J., Amos, W., Barton, D. E. and Fergusonsmith, M. A. (1994). Myotonic dystrophy CTG repeats and the associated insertion/deletion polymorphism in human and primate populations. Hum. Mol. Genet. **3**, 2031-2035.

Rudel, R. and Lehmannhorn, F. (1985). Membrane changes in cells from myotonia patients. Physiological Reviews **65**, 310-356.

Sabourin, L. A., Mahadevan, M. S., Narang, M., Lee, D. S. C., Surh, L. C. and Korneluk, R. G. (1993). Effect of the myotonic dystrophy (DM) mutation on mRNA levels of the DM gene. Nat. Genet. **4**, 233-238.

Sabourin, L. A., Tamai, K. T., Narang, M. A. and Korneluk, R. G. (1997). Overexpression of 3'-untranslated region of the myotonic dystrophy kinase cDNA inhibits myoblast differentiation *in vitro*. J. Biol. Chem. **272**, 29626-29635.

Sakuma, H., Takase, S., Mizuno, Y., Teramura, K. and Hanew, K. (1988). Hypothalamo-pituitary dopaminergic system in patients with myotonic dystrophy. Tohoku Journal of Experimental Medicine **156**, 291-298.

Schein, C. H. (1989). Production of soluble recombinant proteins in bacteria. Bio-Technology **7**, 1141-1147.

Schein, C. H. (1990). Solubility as a function of protein structure and solvent components. Bio-Technology **8**, 308-315.

Seo, H. C., Drivenes, O., Ellingsen, S. and Fjose, A. (1998a). Expression of two zebrafish homologues of the murine *Six3* gene demarcates the initial eye primordia. Mech. Dev. **73**, 45-57.

Seo, H. C., Drivenes, O., Ellingsen, S. and Fjose, A. (1998b). Transient expression of a novel *Six3*-related zebrafish gene during gastrulation and eye formation. Gene **216**, 39-46.

Seo, H. C., Drivenes, O. and Fjose, A. (1998c). A zebrafish *Six4* homologue with early expression in head mesoderm. Biochim. Biophys. Acta **1442**, 427-431.

Serikaku, M. A. and O'Tousa, J. E. (1994). *sine oculis* is a homeobox gene required for drosophila visual system development. Genetics **138**, 1137-1150.

Shaw, D. J., Chaudhary, S., Rundle, S. A., Crow, S., Brook, J. D., Harper, P. S. and Harley, H. G. (1993a). A study of DNA methylation in myotonic dystrophy. J. Med. Genet. **30**, 189-192.

Shaw, D. J., McCurrach, M., Rundle, S. A., Harley, H. G., Crow, S. R., Sohn, R., Thirion, J. P., Hamshere, M. G., Buckler, A. J., Harper, P. S., Housman, D. E. and Brook, J. D. (1993b). Genomic organization and transcriptional units at the myotonic dystrophy locus. Genomics **18**, 673-679.

Spitz, F., Demignon, J., Porteu, A., Kahn, J.-P., Daegelen, D. and Maire, P. (1998). Expression of myogenin during embryogenesis is controlled by Six/*Sine oculis* homeoproteins through a conserved MEF3 binding site. Proc. Natl. Acad. Sci. (USA) **95**, 14220-14225.

Steinbach, P., Glaser, D., Vogel, W., Wolf, M. and Schwemmle, S. (1998). The DMPK gene of severely affected myotonic dystrophy patients is hypermethylated proximal to the largely expanded CTG repeat. Am. J. Hum. Genet. **62**, 278-285.

Steinert, H. (1909). Uber das klinische und anatomische bild des muskelschwunds der myotoniker. Dtsch. Z. Nervenheilkd **37**, 58-104.

Storbeck, C. J., Sabourin, L. A., Waring, J. D. and Korneluk, R. G. (1998). Definition of regulatory sequence elements in the promoter region and the first intron of the myotonic dystrophy protein kinase gene. J. Biol. Chem. **273**, 9139-9147.

Sunahara, R. K., Guan, H. C., O'Dowd, B. F., Seeman, P., Laurier, L. G., Ng, G., George, S. R., Torchia, J., Van Tol, H. H. M. and Niznik, H. B. (1991). Cloning of the gene for a human dopamine D_5 receptor with a higher affinity for dopamine than D_1. Nature **350**, 614-619.

Suzuki-Yagawa, Y., Kawakami, K. and Nagano, K. (1992). Housekeeping Na,K-ATPase α1 subunit gene promoter is composed of multiple *cis* elements to which common and cell type-specific factors bind. Mol. Cell. Biol. **12**, 4046-4055.

Taneja, K. L., McCurrach, M., Schalling, M., Housman, D. and Singer, R. H. (1995). Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues. J. Cell Biol. **128**, 995-1002.

Tassabehji, M., Read, A. P., Newton, V. E., Harris, R., Balling, R., Gruss, P. and Strachan, T. (1992). Waardenburg's syndrome patients have mutations in the human homolog of the *Pax-3* paired box gene. Nature **355**, 635-636.

Thiesen, H. J. and Bach, C. (1990). Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. Nucleic Acids Research **18**, 3203-3209.

Thornton, C. A., Johnson, K. and Moxley, R. T. (1994). Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. Annals Of Neurology **35**, 104-107.

Thornton, C. A., Wymer, J. P., Simmons, Z., McClain, C. and Moxley, R. T. (1997). Expansion of the myotonic dystrophy CTG repeat reduces expression of the flanking *DMAHP* gene. Nat. Genet. **16**, 407-409.

Timchenko, L. T., Miller, J. W., Timchenko, N. A., Devore, D. R., Datar, K. V., Lin, L. J., Roberts, R., Caskey, C. T. and Swanson, M. S. (1996). Identification of a $(CUG)_n$ triplet repeat RNA-binding protein and its expression in myotonic dystrophy. Nucleic Acids Research **24**, 4407-4414.

Ton, C. C. T., Hirvonen, H., Miwa, H., Well, M. M., Monaghan, P., Jordan, T., van Heynigen, V., Hastie, N. D., Meijers-Heijboer, H., Drechsler, M., Royer-Pokora, B., Collins, F., Swaroop, A., Strong, L. C. and Saunders, G. F. (1991). Positional cloning and characterization of a paired box- and homeobox-containing gene from the aniridia reguin. Cell **67**, 1059-1074.

Toy, J., Yang, J. M., Leppert, G. S. and Sundin, O. H. (1998). The *Optx2* homeobox gene is expressed in early precursors of the eye and activates retina-specific genes. Proc. Natl. Acad. Sci. (USA) **95**, 10643-10648.

Treisman, J., Gonczy, P., Vashishtha, M., Harris, E. and Deshplan, C. (1989). A single amino acid can determine the binding specificity of homeodomain proteins. Cell **59**, 553-562.

Treisman, J., Harris, E. and Desplan, C. (1991). The paired box encodes a second DNA-binding domain in the paired homeo domain protein. Genes and Development **5**, 594-604.

Trottier, Y., Lutz, Y., Stevanin, G., Imbert, G., Devys, D., Cancel, G., Saudou, F., Weber, C., David, G., Tora, L., Agid, Y., Brice, A. and Mandel, J. L. (1995). Polyglutamine expansion as a pathological epitope in huntingtons disease and 4 dominant cerebellar ataxias. Nature **378**, 403-406.

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science **249**, 505-510.

Van der Ven, P. F. M., Jansen, G., van Kuppevelt, T. H. M. S. M., Perryman, M. B., Lupa, M., Dunne, P. W., ter Laak, H. J., Jap, P. H. K., Veerkamp, J. H., Epstein, H. F. and Wieringa, B. (1993). Myotonic dystrophy kinase is a component of neuromuscular-junctions. Hum. Mol. Genet. **2**, 1889-1894.

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P. A., Pizutti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F., Eussen, B. E., van Ommen, G. J. B., Blonden, L. A. J., Riggens, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A. and Warren, S. T. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell **65**, 905-914.

Verrijzer, C. P., Alkema, M. J., van Weperen, W. W., Van Leeuwen, H. C., Strating, M. J. J. and van der Vliet, P. C. (1992). The DNA binding specificity of the bipartite POU domain and its subdomains. Embo J. **11**, 4993-5003.

Verrijzer, C. P. and Vandervliet, P. C. (1993). Pou Domain Transcription Factors. Biochim. Biophys. Acta **1173**, 1-21.

Wallis, D. E., Roessler, E., Hehr, U., Telem, D., Rommens, J., Golden, J. A. and Muenke, M. (1998). Missense mutations in the homeodomain of SIX3 are associated with holoprosencephaly. Am. J. Hum. Genet. **63 (Supp)**, A27.

Walsh, F. S., Moore, S. E. and Dickson, J. G. (1988). Expression of membrane antigens in myotonic dystrophy. J. Neurol. Neurosur. Psych. **51**, 136-138.

Wang, J., Pegoraro, E., Menegazzo, E., Gennarelli, M., Hoop, R. C., Angelini, C. and Hoffman, E. P. (1995). Myotonic dystrophy: evidence for a possible dominant-negative RNA mutation. Hum. Mol. Genet. **4**, 599-606.

Wang, Y.-H. and Griffith, J. (1995). Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. Genomics **25**, 570-573.

Watkins, W. S., Bamshad, M. and Jorde, L. B. (1995). Population genetics of trinucleotide repeat polymorphisms. Hum. Mol. Genet. **4**, 1485-1491.

Weintraub, H. (1993). The MyoD family and myogenesis: Redundancy, networks, and thresholds. Cell **75**, 1241-1244.

Wells, R. D. (1996). Molecular basis of genetic instability of triplet repeats. J. Biol. Chem. **271**, 2875-2878.

Werner, M. H. and Burley, S. K. (1997). Architectural transcription factors: Proteins that remodel DNA. Cell **88**, 733-736.

Whiting, E. J., Waring, J. D., Tamai, K., Somerville, M. J., Hincke, M., Staines, W. A., Ikeda, J.-E. and Korneluk, R. G. (1995). Characterization of myotonic dystrophy kinase (DMK) protein in human and rodent muscle and central nervous-tissue. Hum. Mol. Genet. **4**, 1063-1072.

Wilson, D., Sheng, G., Lecuit, T., Dostatni, N. and Desplan, C. (1993). Cooperative dimerization of Paired class homeo domains on DNA. Genes and Development **7**, 2120-2134.

Winchester, C. L. (1997). Expression of myotonic dystrophy candidate proteins. Unpublished PhD, London University, London.

Winchester, C. L., Ferrier, R. K., Sermoni, A., Clark, B. J. and Johnson, K. J. Characterisation of the expression of DMPK and DMAHP in the human eye and implications for pathogenesis in myotonic dystrophy. In Press

Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D. and Pabo, C. O. (1991). Crystal structure of a MATα2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell **67**, 517-528.

Wright, W. E. and Funk, W. D. (1993). CASTing for multicomponent DNA-binding complexes. Trends in Biological Sciences **18**, 77-80.

Xu, P. X., Cheng, J., Epstein, J. A. and Maas, R. L. (1997a). Mouse *Eya* genes are expressed during limb development and encode a transcriptional activation domain. Proc. Natl. Acad. Sci. (USA) **94**, 11974-11979.

Xu, P. X., Woo, I., Her, H., Beier, D. R. and Maas, R. L. (1997b). Mouse *Eya* homologues of the *Drosophila eyes absent* gene require *Pax6* for expression in lens and nasal placode. Development **124**, 219-231.

Zerylnick, C., Torroni, A., Sherman, S. L. and Warren, S. T. (1995). Normal variation at the myotonic dystrophy locus in global human populations. Am. J. Hum. Genet. **56**, 123-130.

Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D. W., Amos, C., Dobyns, W. B., Subramony, S. H., Zoghbi, H. Y. and Lee, C. C. (1997). Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha(1A) voltage dependent calcium channel. Nat. Genet. **15**, 62-69.

Zimmerman, J. E., Bui, Q. T., Steingrimsson, E., Nagle, D. L., Fu, W. L., Genin, A., Spinner, N. B., Copeland, N. G., Jenkins, N. A., Bucan, M. and Bonini, N. M. (1997). Cloning and characterization of two vertebrate homologs of the Drosophila eyes absent gene. Genome Research **7**, 128-141.